Improved Tools for Protein Tertiary Structure Prediction

Shane Steven Sturrock

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh, 1997

Abstract

The most successful method to date for predicting protein tertiary structure from primary sequence data is homology modelling based on alignment with similar sequences of known structure. The use of a variety of computing methods to identify the best similarities is discussed.

Model building based on alignments and the construction of libraries of side-chain conformations is described. The application of sequence alignment modelling to the structure prediction of *Eco*KI type I DNA methyltransferase is shown in the context of corroborative laboratory experiments. Finally, a method is presented which incorporates sequence alignment with secondary structure prediction.

A program — `sss_align` — which incorporates this method was used to make blind fold recognition predictions as part of an international collaborative exercise in the critical assessment of methods of protein structure prediction ('CASP2'). It was shown by this and other assessment methods that `sss_align` will detect similarities between sequences which exhibit as little as 15% identity.

Declaration

I declare that this thesis was composed by myself and the research presented is my own except where otherwise stated.

Shane Steven Sturrock

1997

Acknowledgments

I would like to thank my supervisors, Dr. Andrew Coulson and Dr. John Collins, for their invaluable assistance in completing this work. In addition I am very grateful to Dr. David Dryden and Prof. Noreen Murray for providing an interesting and inspiring problem in the form of the DNA Methyltransferases.

This work was funded by the BBSRC via a CASE award in partnership with MasPar computer corporation, UK, and by a Darwin Trust Scholarship.

I dedicate this work to my fiancee, Katherine Robertson.

Contents

Appendices **attached to the back**

User guide for `sss_align` (code on floppy attached to cover)

Submitted paper: A prediction of the amino acids and structures involved in DNA recognition by type I DNA restriction modification systems

Published paper: Structural modelling of a type I DNA methyltransferase

Chapter 1

Introduction

1.1 Proteins

1.1.1 Building blocks

Proteins are made up of 20 different amino acids, each being chemically distinct allowing great variation in the chemical properties of each protein. All have a central carbon atom ($C\alpha$) in common to which are attached a hydrogen atom, an amino group (NH_2), and a carboxyl group ($COOH$). In addition there is one of 20 different sidechains attached to the $C\alpha$. These amino acids are connected covalently to form a polypeptide chain during protein synthesis by the formation of peptide bonds (Figure 1.1).

The carboxyl group of the first amino acid condenses with the amino group of the next to give a peptide bond, a process which repeats as the chain elongates.

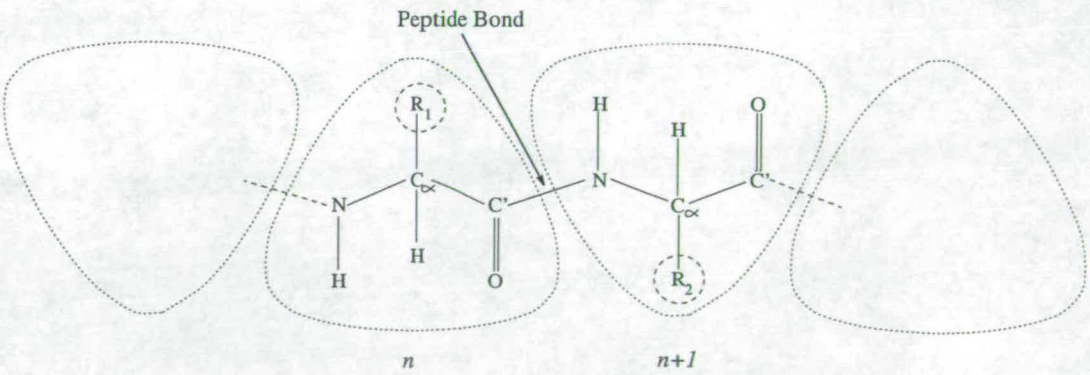


Figure 1.1: *Repeating structure of a polypeptide chain where* R *represents the sidechain and each unit is bounded by the dotted line.*

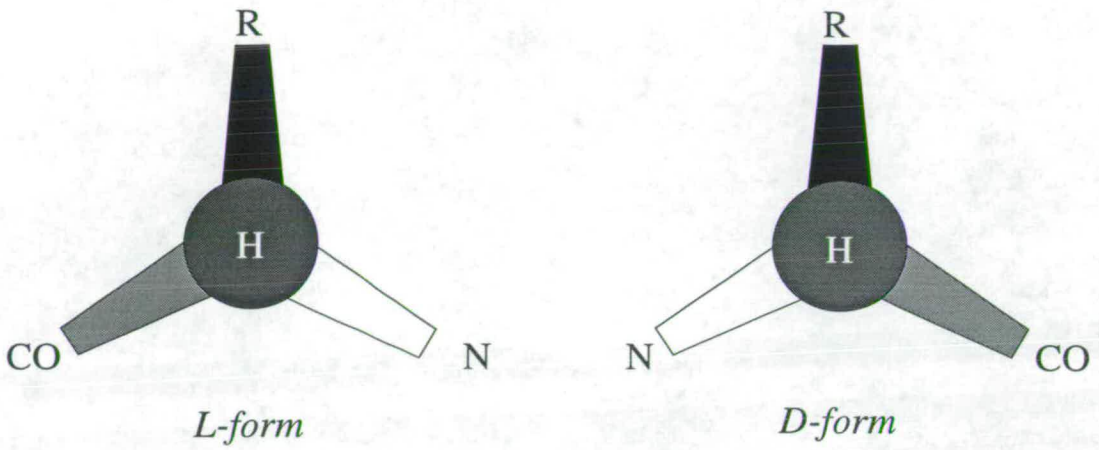


Figure 1.2: *View of L-form and D-form enantiomers shown from above Hydrogen to $C\alpha$ bond.*

The first amino acid in the chain has an amino group that remains intact as does the carboxyl group of the last amino acid and the chain is said to run from its amino terminus (N-terminus) to its carboxy terminus (C-terminus).

Rotation about the bonds in these polypeptides gives the protein a vast range of possible conformations although most chains fold into one energetically preferred conformation. A convention has been adopted to call the angle of rotation around the $N - C\alpha$ bond *phi* (ϕ) and the angle around the $C\alpha - C'$ bond from the same $C\alpha$ atom *psi* (ψ). The presence and position of sidechains builds up a large energy which gives the protein unusual stability. All amino acids (except glycine) are chiral molecules which can exist in two different forms with different *hands*, L- or D-form, (Figure 1.2).

Since biological systems depend on specific detailed recognition of molecules involving differentiation between chiral forms, the translation machinery for protein synthesis has evolved to utilise only the L-form.

Factors which govern the folding are the distribution of polar and non-polar sidechains as well as the proximity of sidechains to each other in the final folded structure. As the protein is synthesised its many hydrophobic sidechains tend to be pushed together in the interior of the molecule to avoid contact with the aqueous environment. Polar sidechains arrange themselves near the outside of the protein where they can interact with the water and other polar groups. Additionally,

peptide bonds are quite polar so they tend to interact with one another and with polar sidechains to form hydrogen bonds. Hydrogen bonding plays an important role in holding the different regions of the polypeptide together in a folded protein.

In addition to hydrogen bonding two cysteine residues in different parts of the polypeptide chain, but adjacent in the three dimensional structure of a protein can be oxidised to form a disulphide bridge. This reaction requires an oxidative environment, and such disulphide bridges are usually not found in intracellular proteins which spend their lifetime in an essentially reductive environment. Disulphide bridges do occur frequently in extracellular proteins that are secreted from cells stabilising the three dimensional structure and they may even hold together different polypeptide chains.

1.1.2 Protein folding

Proteins fold spontaneously into unique conformations even after denaturation which shows that they do not require assistance in reaching their folded state. They are usually compact and globular but may be long and fibrous. The position and chemistry of the different atoms on the surface make each protein specific for binding of both other macromolecular surfaces (as with multimeric proteins made up of more than one individual polypeptide chain) and small molecules.

Creighton (1994) notes that proteins are observed to fold to their final three dimensional structure many orders of magnitude more rapidly than would be expected if folding occurred randomly. The mechanism by which they attain their native conformation is difficult to determine. Historically, emphasis has been placed upon characterising the transient, partly-folded states that appear while folding occurs. However, many of the proposed intermediates and rate determining steps in, for example, the folding of cytochrome *c* may be the result of mis-folded proteins that are not part of the intrinsic folding process. In the absence of such non-productive events which are slowly reversed folding occurs much more rapidly. To study the kinetic processes of protein folding, the protein is first unfolded by exposure to denaturing conditions, e.g. high concentrations of denatu-

rant, extremes of pH, or high temperatures. Under these conditions most unfolded proteins approximate randomly-coiled polypeptide chains, which possess an enormous number of conformations and, consequently, every molecule in a sample of typical size is likely to have a unique conformation at each instant of time, and another unique conformation some 10^{-10} s later. Returning the unfolded protein to less denaturing conditions, its properties can change rapidly and dramatically. In the most extreme case, every molecule might be expected to refold at a unique rate; the only heterogeneity observed in refolding rate is usually that arising from intrinsically slow isomerisations, such as *cis-trans* isomerisation of peptide bonds preceding Proline residues.

Results of the study by Sosnick *et al.* (1994) suggest that the slow steps often seen in protein folding are due to structural mis-organisation which is produced in and stabilised by initial chain condensation. These folding errors are analogous to deep local minima seen in computer simulations. Fast access to the native state depends on avoiding the formation of stable defects which must be corrected for folding to proceed. Kinetically trapped intermediates have been used as evidence for the existence of determinate folding pathways and as the basis for several folding models. These mis-organised intermediates are off the fast pathway(s) and therefore might be considered peripheral to the search for the principals of protein folding. Alternatively, important structural aspects of intermediates detected under slow folding conditions are likely to represent features that do occur on fast folding pathways.

Inside the cell the folding process for many proteins, particularly multidomain structures, is prone to these misfolded species and aggregates. A specialised class of proteins, molecular chaperones, plays an essential role in binding non-native proteins and preventing protein aggregation (Fenton & Horwich, 1997).

Proteins have important roles in the cell as enzymes, metabolic regulators and structural components with each role being defined by the folding of the protein chain. The information required for the folding of each protein is contained in the amino acid sequence and thus, in theory, it should be possible to predict the structure of a protein from sequence alone. Due to the immense nature of such a

calculation (for a protein of 1000 residues, each with three distinct conformations, there are $3^{1000} = 10^{48}$ possible conformations) this is simply not possible by exhaustive evaluation of every conformation. However, the folded conformation as determined by X-ray diffraction or nuclear magnetic resonance (NMR) shows that the three dimensional structures of different proteins, while unique, share several folding patterns. Two patterns are common because they result from repeated hydrogen bonding interactions between the peptide bonds themselves instead of depending on a unique pattern of sidechain interactions. These are known as the β sheet and α helix.

1.2 Protein structure

Proteins seen in nature have evolved to perform specific functions. Functional properties are dependant on their three dimensional structures which arise because particular sequences of amino acids in polypeptide chains fold to generate, from linear chains (primary structure), compact domains with specific three dimensional structures (tertiary structure) which can in turn join with other polypeptide chains to create large multimers (quaternary structures). The three dimensional structure brings together the various amino acids that form the functional region or active site.

1.2.1 α helix

A α-helix results when a single polypeptide chain turns regularly about itself to make a rigid cylinder in which each peptide bond is hydrogen bonded to other peptide bonds elsewhere in the chain, (Figure 1.3). They are found when a stretch of consecutive residues all have the ϕ, ψ angle pair approximately -57^o and -47^o. There are 3.6 residues per turn of helix with hydrogen bonds between $C' = O$ of residue n and NH of residue $n+4$ (see figure). Thus all NH and $C' = O$ groups are joined with hydrogen bonds except the first NH groups and the last $C' = O$ groups at the ends of the helix.

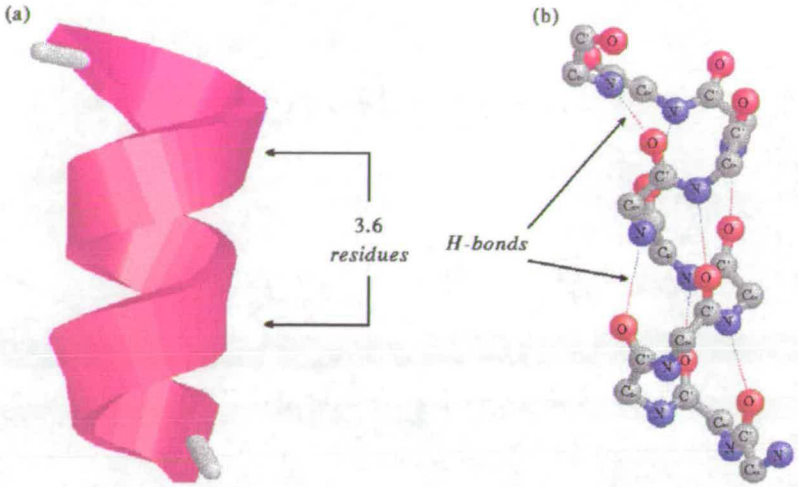


Figure 1.3: *The α helix.* (a) *Idealised diagram of the path of the main chain.* (b) *Ball and stick model of α carbon backbone showing position of helix stabilising hydrogen bonds (dotted lines).*

Theoretically, the helix can be right- or left-handed depending on the screw directions of the chain but the right-handed form vastly outnumbers the left-handed because for L-form amino acids the close approach of the side chains and the $C' = O$ group is not favourable.

Different sidechains have been found to have weak but definite preferences either for or against being in α helices. ala (A), glu (E), leu (L), and met (M) are good α helix formers, while pro (P), gly (G), tyr (Y), and ser (S) are very poor.

These preferences have been used in attempts to predict the secondary structure (Chou & Fasman, 1974ab, Garnier, Osguthorpe & Robson, 1978, Rost & Sander, 1993) but have not proved strong enough to give reliable predictions as they are about 70% accurate.

Helices are commonly located along the outside of the protein with one side facing the solution and the other toward the hydrophobic interior. With 3.6 residues per turn there is a tendency for the sidechains to change from hydrophilic to hydrophobic with a periodicity of three to four residues.

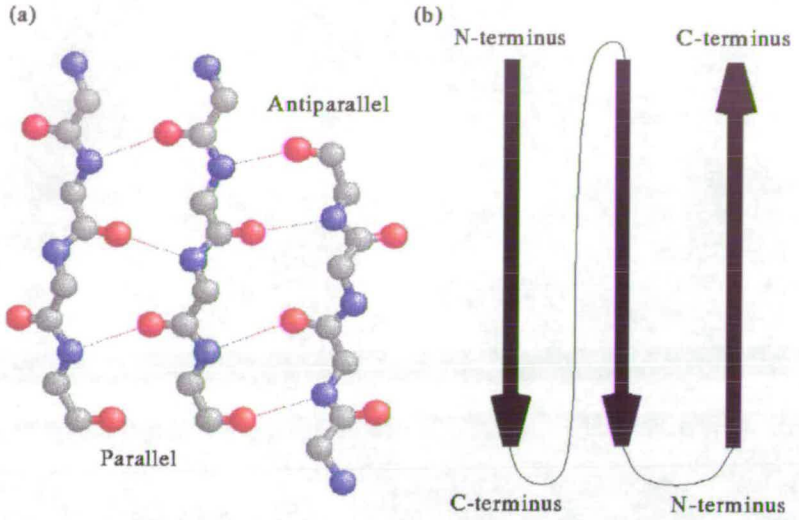


Figure 1.4: *Parallel and Antiparallel β sheet showing* (a) *the arrangement of the polypeptide chains joined by hydrogen bonds (dashed lines) and* (b) *a diagrammatic representation of the biochemical direction of the same sequence/structure.*

1.2.2 β sheet

The β sheet makes up extensive reqions of the core of many globular proteins. It consists of an extended polypeptide chain which folds back and forth upon itself with each chain either running in the direction opposite to its immediate neighbours or running parallel with each strand joined by a large loop, helix or an whole domain (Figure 1.4).

The β strands are aligned adjacent to each other such that hydrogen bonds can form between $C' = O$ groups of one strand and NH groups of another. β sheets can be made up of parallel, anti-parallel or mixed sheets although there is a strong bias against mixed β sheets with only about 20% of the strands in known structures being of this mixed type. Almost all β sheets have their strands twisted, always right-handed.

1.2.3 Loop regions

The structures of proteins are combinations of secondary structure elements, helices or sheets, and these are connected by loop regions of various lengths and

irregular shape. Secondary structure elements form the hydrophobic core with the loop regions at the surface of the molecule and do not form hydrogen bonds to each other but can hydrogen bond with the water molecules of the solvent.

Homologous amino acid sequences from different species exhibit insertions and deletions of a few residues almost exclusively in the loop regions.

1.2.4 Structure determination

Various techniques are used to determine the levels of structure within proteins. X-ray diffraction involves purification of a quantity of protein. Crystals that will diffract X-rays strongly, are of a suitable size and are stable in an X-ray beam, are grown. Several isomorphous derivatives containing heavy atoms are prepared and the diffraction patterns from directing the X-ray beam onto a regular, repeating array of many identical molecules is recorded. From the intensities and phases of the best data the electron density needs to be solved at low resolution, then cyclic comparison and computation of the structure with more data is used to refine the structure to obtain the electron density distribution at high resolution. Knowledge of standard bond lengths and bond angles, along with the primary sequence of the protein are used to fit the atoms to this distribution, represented as a 3D contour map and a model is built.

NMR has an advantage since the protein does not need to be crystallised which may distort soluble structures, and hydrogens do not show up at all with X-ray crystallography so their positions have to be inferred. Instead the protein is left in solution and placed in a uniform magnetic field. The purified sample is subjected to a short pulse of electromagnetic energy whose frequency is centred within the spectral region of interest where it excites all of the nuclei of interest (e.g. all ^{1}H or all ^{13}C). The transitory magnetisation of the sample, produced at right angles to the static magnetic field, generates a signal in the receiver coil which is detected in the audio frequency range as an offset from the radio frequency of the transmitter. This signal is digitised by an analogue to digital converter and stored

in a computer. Fourier transformation of this signal yields the NMR spectrum from which the structure can be deduced.

1.3 Structure prediction

1.3.1 Introduction

Predicting the three dimensional structure of a protein from its amino acid sequence is the major unsolved problem in structural biology. Theoretically, a computer program that could simulate the action of the simple physical laws that operate in a test tube, or a living cell, when a polypeptide chain with a specific amino acid sequence folds into a precise three dimensional structure is needed. Unfortunately, the complexity of the task of searching through all possible conformations of the chain to find those with low energy requires an enormous amount of computing time ($\approx 3^2 n$ conformations where n is the number of residues in the sequence). Accepting this as an impossible task, other means are required; comparison of new sequences to those already known is such a method.

1.3.2 Sequence comparison

A basic and very useful premise for the prediction of protein structure is that a similar protein sequence is likely to share a similar structure and function since high similarity implies evolutionary relatedness. Identifying a sequence as similar to an unknown query is complicated by two factors as a result of mutation:

- Residues can be replaced with other similar residues which have little effect on the chemical and structural properties of the protein.

- In loop regions it is common for insertion or deletion of residues to occur over time.

Taking account of these it is possible to search databases for close or even quite distant relatives of the query sequence.

Alanine	A	Leucine	L
Arginine	R	Lysine	K
Asparagine	N	Methionine	M
Aspartate	D	Phenylalanine	F
Cysteine	C	Proline	P
Glutamate	E	Serine	S
Glutamine	Q	Threonine	T
Glycine	G	Tryptophan	W
Histidine	H	Tyrosine	Y
Isoleucine	I	Valine	V
Note: The following codes are used in cases of uncertainty.			
Unknown	X	E or Q	Z
N or D	B		

Table 1.1: *Single letter codes for each amino acid plus the three ambiguous codes.*

1.3.3 Alignment

Ignoring mutational substitution matrices for the moment it is important to understand how an alignment is made.

In applying computers to protein searches it is sensible to use the single-letter amino acid codes (Table 1.1) instead of the usual three-letter codes common in biochemistry.

An alignment is a means of showing how two sequences may be related via a number of modifications of the sequences. Events such as the replacement of a residue with another having similar properties will show up as will insertion or deletion events which appear as gaps in one sequence or another. In order to retrieve such alignments a mathematically optimal method should be used.

1.3.4 Dynamic programming

Similarity searches require a solution to one of three types of alignment problem (Collins & Coulson, 1987):

- Type I — Find the best end to end alignment of two finite sequences.
- Type II — Find the region of an indefinitely long sequence most similar to a short query.
- Type III — Find the most similar pair of subsequences from two indefinitely long sequences.

The method of alignment first applied to biological sequence analysis by Needleman & Wunsch (1970) used dynamic programming techniques to track the best paths through a match matrix. This two dimensional array represented all possible pair combinations that could be formed from the two sequences with each combination being assigned a suitable value for matches and mismatches. For a 1000×1000 match matrix there are approximately 10^{600} possible paths (Waterman, 1988) so direct searching of all possible paths by, for example, recursion would be impossible. Instead, the method involved a technique referred to as backtracking where each point in the matrix becomes the sum of the scores to reach that point. Tracing the best path in such a matrix only involves finding the highest point and then stepping diagonally down to the next highest value until a score of 0 or the edge of the array is reached at which time the alignment is complete. Sideways or vertical steps indicate gaps in the alignment which correspond to an insertion or deletion from one sequence or the other and incur a penalty to limit propagation of such gaps.

Smith & Waterman (1981) produced a modified algorithm where the score assigned is derived from a weighting scheme such as that produced by Dayhoff (1978). The best path through the matrix is obtained (type III or best local similarity) using backtracking as with the Needleman Wunsch method. An additional modification of the method has also been added (Gotoh, 1982) where the gap penalty is no longer linear. Here gaps cost more to start than to extend and these

are referred to as 'affine' gaps. The reasoning behind these was to emulate the way gaps in real proteins form where each missed residue is not a single event but the whole section is removed in one piece thus it should be treated as a single event. However, this method does have problems with shorter sequences with single residues deleted as the cost of opening a single gap is higher than with normal linear non-affine gaps and some potentially similar sequence alignments may be overlooked. Thus it is necessary to use both affine and non-affine searches in order to optimise the set of results.

The Smith & Waterman algorithm was the chosen method of Coulson *et al.* (1987) and is considered the most useful method for current sequence analysis problems in biology (Waterman, 1988) producing high resolution alignments that are guaranteed to optimise a well-understood alignment score.

1.3.5 Alternative methods

Increasing database size has resulted in slower searches due to the linear relationship between database size and the time taken for a search to be completed. The `FASTA` program (Pearson & Lipman, 1988) is one of the most widely used sequence comparison programs because it is more rapid than the exhaustive dynamic programming algorithms. The enhancement in speed is the result of examining only regions of strong identity, thus reducing greatly the number of comparisons required and speeding up the search by more than a factor of one hundred. `FASTA` is a two pass method where the initial high speed run screens out unlikely alignments and a second exhaustive search is performed on the remaining subset of the database. Unfortunately this often results in interesting alignments down in the 25–30% identity region being missed.

`BLAST` (Altschul *et al.* 1990) is another high speed alternative to dynamic programming which has gained wide acceptance as a searching tool. `BLAST` is a local similarity method capable of finding exact or close ungapped matches very quickly, the scanning phase being based on a determinisitic finite automaton or finite state machine. Hits are extended to find locally maximal segment pairs and multi-

ple segment pairs per sequence pair compared exceeding some cut-off score are reported which compensates for the lack of gaps in many cases.

Other searching schemes based on alternative methods appear regularly and are useful for certain problems but compromises must be accepted to gain speed on slow machines. An alternative is to use a faster machine and implementations of the Smith & Waterman method have appeared on parallel architectures to great effect (Coulson *et al.* 1987).

1.3.6 Replacement tables

When comparing sequences of amino acids it is necessary to account for conservative replacements which will result in a similar structure although the sequence of residues differs. As a measure of distance or similarity it is necessary to use a *weighting* scheme (Collins & Coulson, 1987):

• Distance scores have a match weighting set to zero. This means a perfect match will have a zero score and all imperfect matches will exhibit a negative score. Where the value for mismatches and insertions/deletions (indels) are given a unit weighting, the score will equal the number of changes needed to convert one sequence into the other. This is often referred to as the evolutionary distance. In the biological context, it does not imply that an evolutionary process generated the sequences being compared from an ancestral sequence.

• Similarity scores have positive matches and negative mismatches. They tend to favour paths with many matches. The cost of including a gap must always be more negative than the cost of a mismatch otherwise no mismatches will be reported. Instead, adjacent indels on opposite strands will be found.

One of the most widely used similarity tables is the mutation data matrix (MDM) developed by Dayhoff and colleagues (Dayhoff *et al.* 1978). The first MDM, expressed as a log odds table, was derived from over 400 accepted point mutations (evolutionary replacements of one amino acid for another at homologous positions) between present-day sequences and inferred ancestral sequences. The

relative frequency of exposure of each type of amino acid to mutational change and relative mutability were also taken into account. This system gives added weight (i.e. higher scores) to identities between amino acids that are relatively rare in proteins (e.g. cysteine) compared to identities among common amino acids. Additionally, replacements which have occurred frequently in evolution (such as methionine to leucine) still receive a positive score while unlikely substitutions receive negative scores. A set of suitable values was derived by comparing various closely related proteins and determining the probability that a given amino acid would be replaced by any other in a fixed time. Dayhoff also described how the similarity coefficients varied with the evolutionary distance measured in PAMs (accepted point mutations) where 1 PAM corresponds to the appearance of one substituted amino acid residue in a pair of related proteins, per 100 amino acids aligned. This allows comparisons to be tailored for any particular evolutionary distance.

Over the years a number of researchers have modified the PAM tables or created their own which are derived from more recent data in order to improve sensitivity for example BLOSUM (BLOcks SUbstitution Matrix) tables (Henikoff & Henikoff, 1992).

1.4 Statistical significance of protein sequence similarities

A database search will produce a list of scores for the similarity or distance between the query sequence and each database entry. Assessing which scores are significant is an ongoing problem. Significance is usually expressed as the number of times that a particular score would occur by chance for that particular database. A significant result is one where this number (expected frequency) is less that 1. For example, a result of 0.01 implies that such a result would only occur "by chance" in a search of a database 100 times larger than that actually used (Collins & Coulson, 1990).

In order to calculate this number, the entire set of results can be stored as an histogram and fitted to a Poisson distribution.

1.5 Approaches to structure prediction

While the standard sequence alignment methods have proved valuable in detecting sequence similarities they are all essentially one dimensional comparisons. Simply put, all the methods are based on residue by residue comparison and do not take any account of the position within the structure. The first level of extra information is secondary structure and various schemes for prediction have been tried.

1.5.1 Secondary structure prediction

If the homology search fails to reveal any sequence homology with a protein of known tertiary structure it is still possible to predict secondary structure by relying on which amino acids are commonly found in α helices and which in β sheets (Chou & Fasman, 1974ab). However, such predictions do not have a high degree of confidence with the possible exception of transmembrane helices. Rost & Sander (1993) have used neural networks as a means of improving predictions but are still only about 70% accurate, less so for predominantly β strand proteins.

Results of secondary structure prediction attempts suggest that about 60% (Garnier *et al.* 1978) of the secondary structure is determined by local interactions but global tertiary structure imposes local secondary structure in some regions, thus the ability of a sequence to form helix, sheet or loop structures is dependent not only on the sequence of that region but also on the environment in the tertiary structure. It has been shown that identical sequences of up to five residues in length can form both α helices in one structure and β sheets or a loop in another (Sippl, 1990).

While the accuracy of secondary structure prediction is low a large fraction of errors occur at the ends of α helices or β strands. The central regions of these secondary structure elements are often correctly predicted but methods do not always distinguish between α helices and β strands. Unfortunately, incorrect secondary structure predictions can hinder the ability to use these as a basis for tertiary structure prediction. For instance Russell *et al.* (1996) uses predictions to align against a database of secondary structures by reducing the secondary structure elements to single letters and perform a Smith & Waterman search with these simple sequences. If a predicted helix or strand is incorrect the ability to align the sequences is affected. However, good results have been obtained with this method where sequences with no similarity have been identified as having the same structure.

1.5.2 Common structures

Many completely different amino acid sequences give similar three dimensional structures. It is estimated that there are fewer than 1000 (Chothia, 1992) topologically different domain structures, so far about 400 different domain structures have been observed.

Proteins with homologous amino acid sequences have similar three dimensional structures. Usually, they also have similar functions although there are some exceptions known where genes for ancient enzymes have been recruited at a later stage in evolution to produce proteins with quite different functions. Once a novel gene has been cloned and sequenced, a search for amino acid sequence homology between the corresponding protein and other known protein sequences should be made. Usually, this is done by comparison with databases of known protein sequences using one of the standard sequence alignment computer programs as previously described.

1.5.3 Structural motifs

As has already been stated, the final structure of a protein, while unique, is made up of subunits which recur throughout many varied proteins. Observation of the frequency of occurence of particular residues in particular secondary structures has shown that composition plays an important role in determining the secondary structural elements (Chou & Fasman, 1974ab, Garnier *et al.* 1978, Rost & Sander, 1993). This results in a marked similarity of sequences which share similar structure and these are referred to as structural motifs. Essentially these are small building blocks from which full protein structures are made. Being able to recognise motifs allows secondary structure prediction for example if a motif can be identified in many sequences and that motif corresponds to a helix (bearing in mind the work of Sippl) then there is a good chance that the regions that are the same as or similar to the motif are also helices. Searching a database of structural motifs with an unknown sequence may identify the secondary structure elements and, in addition, drastically cut down the work needed to predict a full tertiary structure because the subunits have limited freedom of movement and as a result reduce the range of possible conformations for the polypeptide chain. The problem at the moment is recognising which sequences of residues will fit into a particular structure and present scoring is inadequate, typically only recognising sequences which are 30%+ identical to the motif. Jones & Thornton (1993) note that sequences with less than 5–10% identity can still exhibit high structural similarity.

1.5.4 Profiles

In attempting to improve the sensitivity of alignments new methods have tried to include data from secondary and tertiary structure, either real or predicted. One of the simplest methods are profiles (Gribskov *et al.* 1987). Profiles include positional information within a scoring table by having a suitable set of scores for every position in the probe sequence rather than the traditional 23×23 matrix. This table takes the form of a list which is $n \times 23$ long where n is the number

of residues in the query sequence and 23 is the number of letters used to identify all the 20 residues and any ambiguous residues. In addition to the scores there can also be a varying gap penalty which allows the method to place gaps in less structurally defined regions such as loops. This information can be derived from simple secondary structure prediction or on multiple alignments of a family to produce a hybrid consensus sequence which ideally should be the sequence most similar to all the sequences of the family aligned (not necessarily a real sequence).

Once a full three dimensional structure for the protein is known reliablity of alignments based on actual secondary structure increases but now the added dimension can be used to adjust scoring dependent on the actual position of residues in the structure. Position dependant scoring tables have been significantly more sucessful than traditional scoring schemes in some cases but are still dependant on where the values in the scheme come from. The original profile method took a multiple sequence alignment and for each position found a residue which was most similar to all those at that position and used this to build a consensus sequence. Statistical modification of the PAM table entry for that residue versus every other residue took into account the variablity of that position in the multiple sequence alignment and resulted in a table which should detect other members of the same family from which the table was created. Additionally, the profile contains a gap penalty for each position which is derived from observed gap frequencies in the multiple sequence alignment and should reflect the secondary structure because with a true family gaps are most likely in regions of loop. Lowering the cost of gaps in loop regions is a way of including secondary structure in a one dimensional search. It is possible to force specific gap costs if the true secondary structure for at least one of the probe sequences is known. Gribskov (1990) notes that any set of properties could be represented as similarity or difference scores for pairs of amino acids such as hydrophobicity, structural preference or sidechain volume in addition to the modified PAM scores he used.

1.5.5 Protein fold recognition

Basic sequence analysis runs into a grey area at about 25% identity where alignments could be correct or incorrect but the noise floor obscures this. Various methods have been tried to improve sensitivity.

Ponder & Richards (1987) attempted to find sequences which would be compatible with a given backbone. The evaluation was based on a simple van der Waals potential and so models were effectively scored on the degree of overlap between sidechain atoms. Additionally, the core was required to be well packed which was achieved by considering the conservation of sidechain volume. In order to fit the sidechains of a given sequence into the backbone, an exhaustive search was made through a *rotamer library* of sidechain conformations. If after searching rotamer space the sidechains could not be fitted successfully into the protein core, then the sequence was deemed incompatible with the given fold. Sensitivity of this method was low because without allowing for backbone shifts, the packing requirement of a given backbone was far too specific. Only sequences very similar to the native sequence could be fitted to the fixed backbone.

Eisenberg *et al.* (1992) translated the sequence of positions in a structure into environments which describe the region of three dimensional structure in relation to its surroundings. In order for a residue to fit this environment it must be compatible with it. In theory this allows detection of much more distant homologues.

Sippl (1990) traced the favoured atomic interactions for each position in a peptide chain. By knowing what sort of atomic distances each residue type will prefer from others it is possible to build up a pseudo-energy term by simple alignment methods, the lowest energy corresponding to the best compatibility of a sequence to a structure. The potential method has come to be known as fold-threading, in essence it ignores sequence and concentrates on the fold. However, fold-threading has yet to convincingly demonstrate an ability to reliably detect similarities which do not have some sequence similarity so if it were possible to improve the sensitivity of sequence similarity alignment the same alignments should be arrived at.

Taylor & Orengo (1989) described a method of aligning three dimensional structures using an extension of the dynamic programming algorithm. In order to calculate the optimal path for an alignment allowing for gaps etc. a second lower level of dynamic alignment is added, hence this method has become known as the double dynamic programming algorithm. For every position in the traditional match matrix a second matrix of equal size is computed where the scores entered consist of distances of all residues in the structure centred on the two residues from the upper level. The optimal path through the lower level is computed and the highest score returned to the upper level. This process continues until all elements of the upper layer are filled and then the optimal path through that matrix is calculated to give the alignment (Figure 1.5).

The complexity of this method is greater than with the traditional dynamic algorithm. Jones *et al.* (1992) extended the double dynamic algorithm to allow alignment of structures with sequence. This method requires the ability to match a given sequence with real co-ordinates of a structure using Sippl potentials to score each residue with each position in the structure. By threading a sequence segment onto templates of known conformations (Figure 1.6) it is possible to evaluate the probability of that sequence folding into that structure. This method has proved successful at detecting similarities in structures where little sequence identity exists.

Wilmanns & Eisenberg (1992) extended the profile method by representing the positions as environments defined by polarity, area buried of its sidechain and its secondary structure while also including residue pair preferences as described by Sippl (1990) while Lüthy *et al.* (1992) used 3D profiles as a means of assessing whether a model protein was compatible with its sequence. For models where a region is incorrect this can be detected by using a moving window on the structure assessing part of the structure at a time.

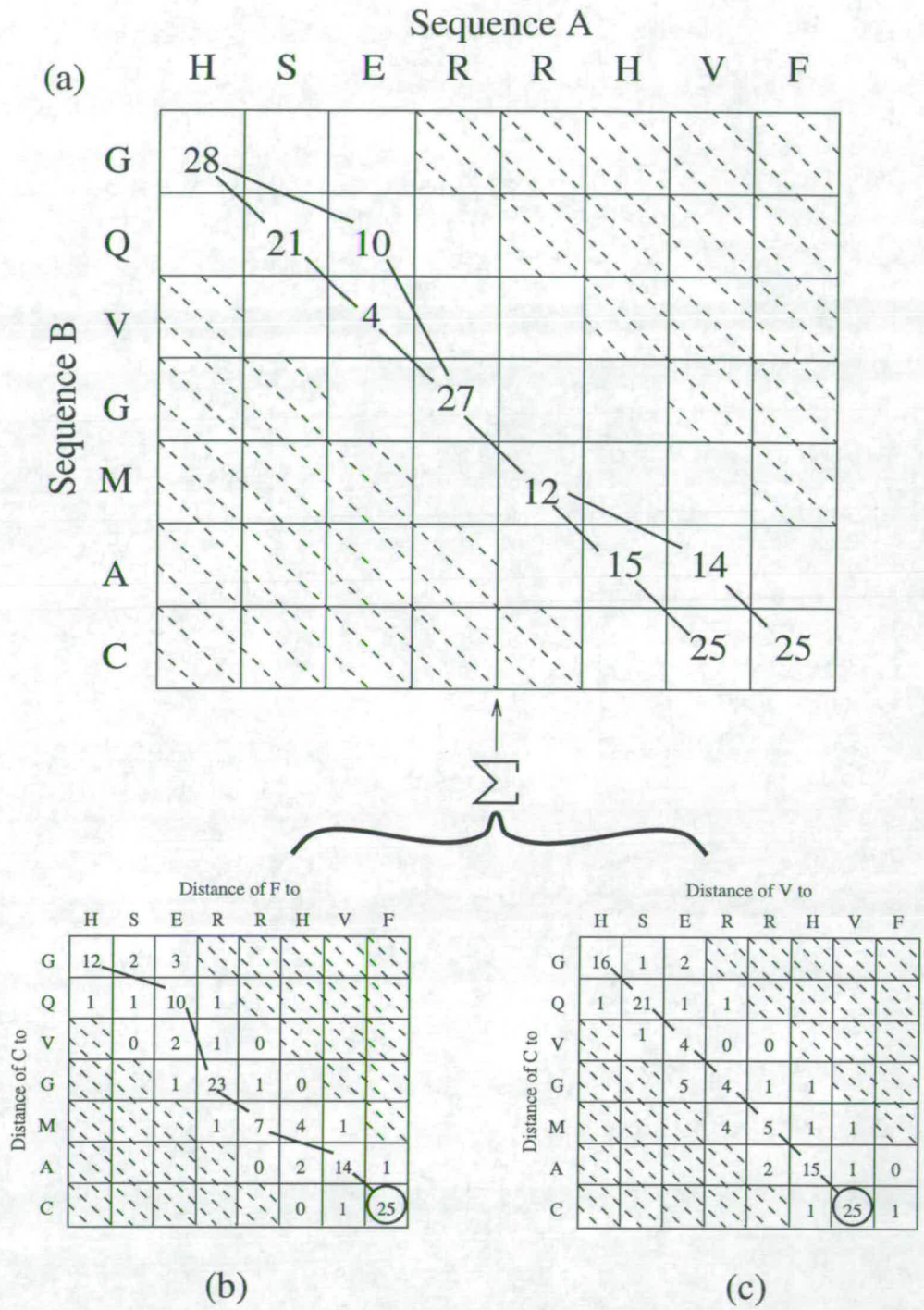


Figure 1.5: *Double dynamic programming algorithm.* (a) *is the score matrix between 2 peptide sequences.* (b) *is the score matrix for comparison of all distances centred on residue C in sequence B and residue F in sequence A. Values from the best path are accumulated in the upper level as are those for* (c). *Note that the values carried into the upper level are added to any already there. Finally a dynamic algorithm is used to calculate the best path through the upper matrix.*

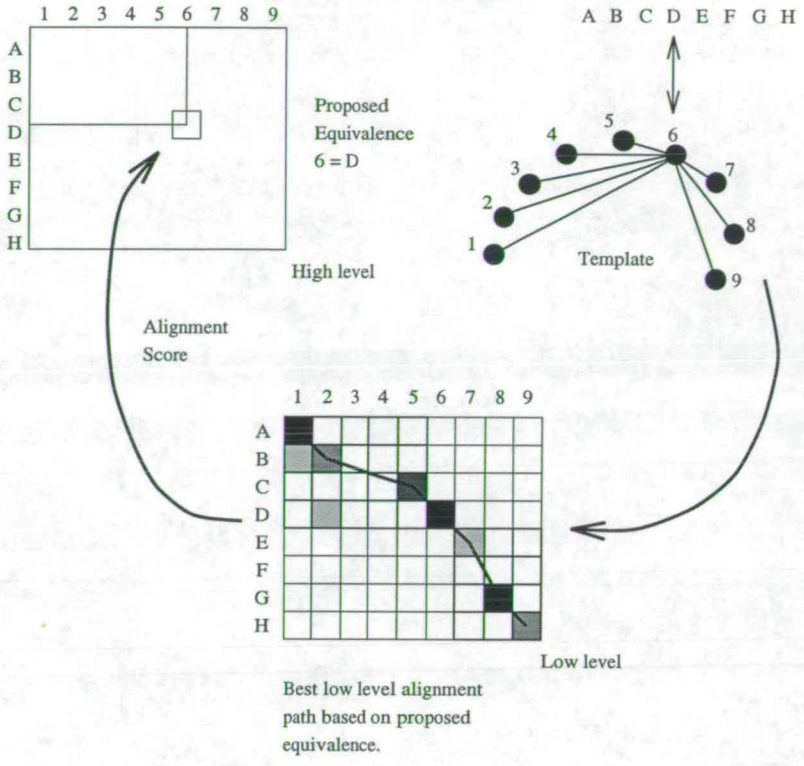


Figure 1.6: *Double dynamic algorithm applied to threading a sequence onto a fold.*

1.5.6 Rotamers

Rotamers are a means of describing a limited set of conformations for each sidechain. Holm & Sanders (1991) performed Monte Carlo optimisation with simulated annealing and precalculation of all possible rotamer pairwise interaction energies. Eisenmenger *et al.* (1993) evaluated rotamers as a means of configuring sidechains and found that where sidechains were placed on the true backbone for that sequence the results were very good with over 50% of sidechains correctly predicted. However, using the backbone of a close homologue (60% identity) resulted in poorer performance with an error rate 40% higher indicating a similar problem to that with the Ponder and Richards template method in that lack of flexibility in the backbone results in poor performance. Even so, the predictions are a good place to start energy minimisation.

1.5.7 Energy minimisation

A physical approach to the prediction of native three dimensional structure in its pure form based on the assumption that the native conformation corresponds to the structure with the lowest free energy and is thus in a state of thermodynamic equilibrium. A view backed by *in vitro* observations that many proteins refold successfully after denaturation.

To solve the problem of folding a protein we should be capable of calculating all essential terms of the free energy of a trial protein conformation with an accuracy sufficient to ensure the uniqueness of the native conformation. Also, a procedure to locate the global minimum of the energy function in the giant space of conformational possibilities using a limited set of function evaluations is needed. Neither of these problems is close to a solution. The free energy of a protein consists of a potential energy in vacuo, the free energy of solvation and a term proportional to the conformation entropy of the polypeptide. In turn, the solvation energy consists of the electrostatic free energy and surface free energy related to hydrophobicity. The difficulty posed by calculating these energies compared to a simple vacuum potential energy and its derivatives has led to the opinion that the true energy of a protein is so complicated that we have neither the means nor hope of determining it properly. The inability of the potential energy in vacuo to replace the true free energy in structure prediction calculations has led to attempts to design functions discriminating between native conformations and incorrect models (false positives). Typically these functions are related to the missing solvation free energy term or are derived from statistics of atomic contacts or accessibilities in the database of known three dimensional structures. Unfortunately such functions do not guarantee that there are no false positives simply because it is not a true free energy calculation.

When applying energy minimisation algorithms one searches for a minimum energy configuration of a system by moving along the gradient of the potential energy through configuration space. Since in this way one basically moves only downhill over the energy hypersurface, energy minimisation yields only a local

minimum energy configuration which is generally not far from the initial one. This is not a problem when the structure being minimised is close to the actual minimum, but when starting from a random coil methods of getting over the kinetic barriers to find the true global minimum must be used.

1.6 Homology modelling

After using a variety of sequence alignment methods a sequence with known structure may be identified which is similar to the unknown and has a known tertiary structure in the database. It then becomes possible to build a model by mutating the known structure sequence to the unknown structure sequence and then minimising the energy of that structure to give a possible prediction.

A three dimensional model of the novel protein can be constructed in a computer display on the basis of the sequence alignment and the known three dimensional structure. This model can then serve as a basis for identifying amino acid residues involved in the active site or in antigenic epitopes, and the model can be used for protein engineering, drug design, or immunological studies.

As an example here is the process of searching an unknown sequence right up to a final model based on homology. The sequence used is human angiogenin which induces vascularisation of normal or malignant tissues abolishing protein synthesis by specifically hydrolysing cellular tRNAs. Within the annotations for this sequence in SwissProt a note is made of its similarity to the pancreatic ribonuclease family. Therefore, when a search is made using the Smith & Waterman method a large number of ribonuclease hits appear which is not surprising. Fortunately, there is a structure known for one example and it is from this that a structure for angiogenin can be built. The alignment (Figure 1.7) is good although there are gaps and regions of low identity.

Looking through the annotations for the bovine ribonuclease shows that the active site is made up of three residues, histidine 38, lysine 67 and histidine 145. Checking the alignment shows that all three are preserved in the alignment. In

```
ID   RNP_BOVIN      STANDARD;      PRT;   150 AA.
AC   P00656;
DE   RIBONUCLEASE PANCREATIC PRECURSOR  (EC 3.1.27.5) (RNASE 1) (RNASE A).
OS   BOS TAURUS (BOVINE), AND BISON BISON (AMERICAN BISON).
RA   CARSANA A., CONFALONE E., PALMIERI M., LIBONATI M., FURIA A.;

  SEQ IDENTITY 34.04%;  SEQ CONSERVATION 63.12%;

  Matches    48;  Conservative   41;  Mismatches  52;  Indels   8;  Gaps   6;

Db       3 LKSLVLLSLLVLVLLLVRVQPSLGKETAA-AKFERQHMDSSTSAASSSNYCNQMMKSRNL 61
           :::| :| |||:|| |  ..|:|:::::  : |  || |:. : : ...||: :|: |.|
Qy       2 VMGLGVL-LLVFVLGLGLTPPTLAQDNSRYTHFLTQHYDAK-PQGRDDRYCESIMRRRGL 59

Db      62 TKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNC 121
           |.  || :|||:|.:  .::|:|.:|| .  ..: .  :|  :. :|.|:  |:| .| |
Qy      60 TSP-CKDINTFIHGNKRSIKAICENKNGNPHRENLRISKS--SFQVTTCKLHGGSPWPPC 116

Db     122 AYKTTQANKHIIVACEGNPYVPVHFDASV 150
            |::| : ::::||||..  :|||:| |:
Qy     117 QYRATAGFRNVVVACENG--LPVHLDQSI 143
```

Figure 1.7: *Alignment of bovine ribonuclease* (Db) *with human angiogenin* (Qy).

addition the regions of gaps and poor identity are loops while helix and sheet structures are more similar. The process of modelling begins with loading the known structure into a molecular modelling package; in this case Tripos `sybyl` was used. The PDB entry from the Brookhaven database was used. This entry lacks the signal sequence of 26 residues which corresponds well to the signal in angiogenin which is 24 residues long; note the two gaps in the alignment in this region. Now that the start of the structure is known the ribonuclease sequence can be progressively mutated into the angiogenin sequence so that the replacement residues take up the same conformation as those in the ribonuclease. Where gaps occur in the alignment residues are either excised or inserted depending on which sequence contains the gaps. The final state is to energy minimise the structure which now corresponds to the angiogenin sequence. A few obvious clashes of residues can be tweaked out by hand to get clear of the active site. The result is a structure which has a low energy configuration and is very likely to be close to that of the actual protein (Figure 1.8). This is an excellent example of homology being indicative of function and structure.

Comparison of the model with the known structure (Acharya *et al.* 1994) shows that the prediction of loops and insertions or deletions from the template structure introduced errors as can be seen in Figure 1.8(d) indicated by lighter colours.

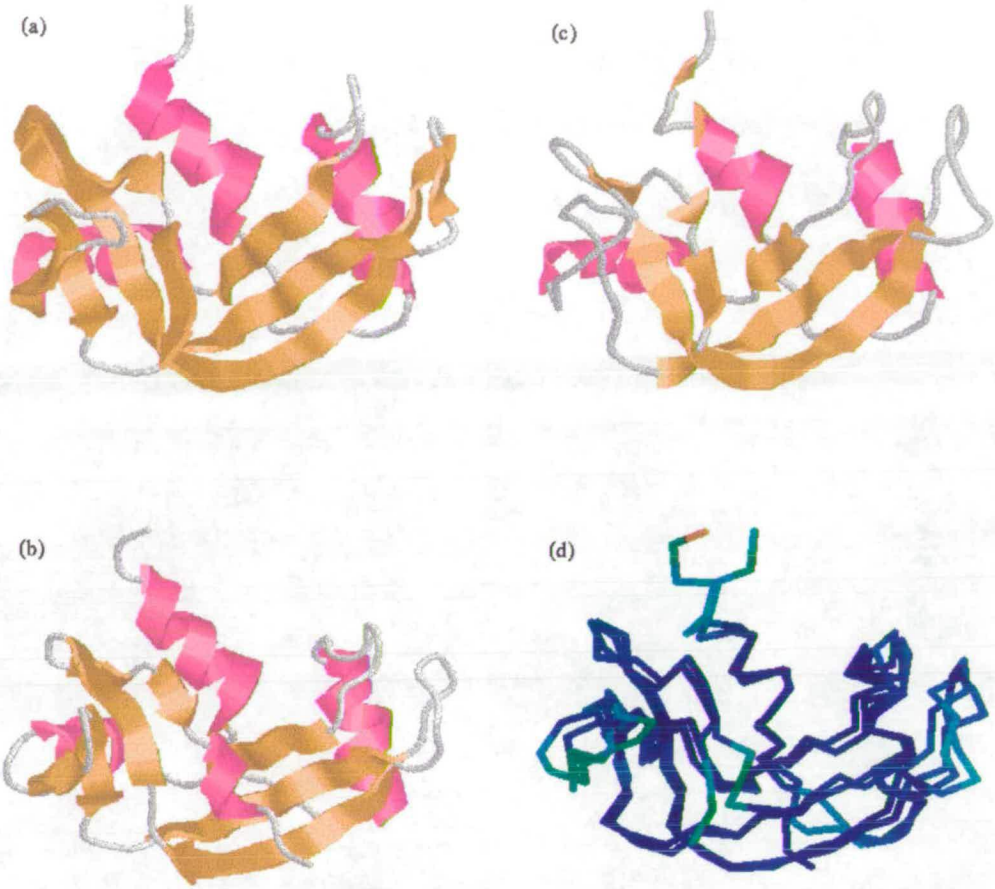


Figure 1.8: (a) *ribonuclease,* (b) *angiogenin,* (c) *predicted angiogenin based on ribonuclease using sequence alignment in Figure 1.7,* (d) *Overlaid backbones of predicted and real angiogenin structures (RMS Error: 0.26 Å).*

However, examination of the active site shows that the model is very close with the correct position and orientation for all three residues (Figure 1.9).

1.7 Summary

High performance computers are beginning to allow biologists to make attempts at problems which were previously unimaginable as can be seen with tertiary structure prediction attempts. Massively parallel computing has been successfully applied to the exhaustive one dimensional Smith & Waterman algorithm in a number of cases and building on this work to allow rapid prediction of protein structure is an obvious extension.

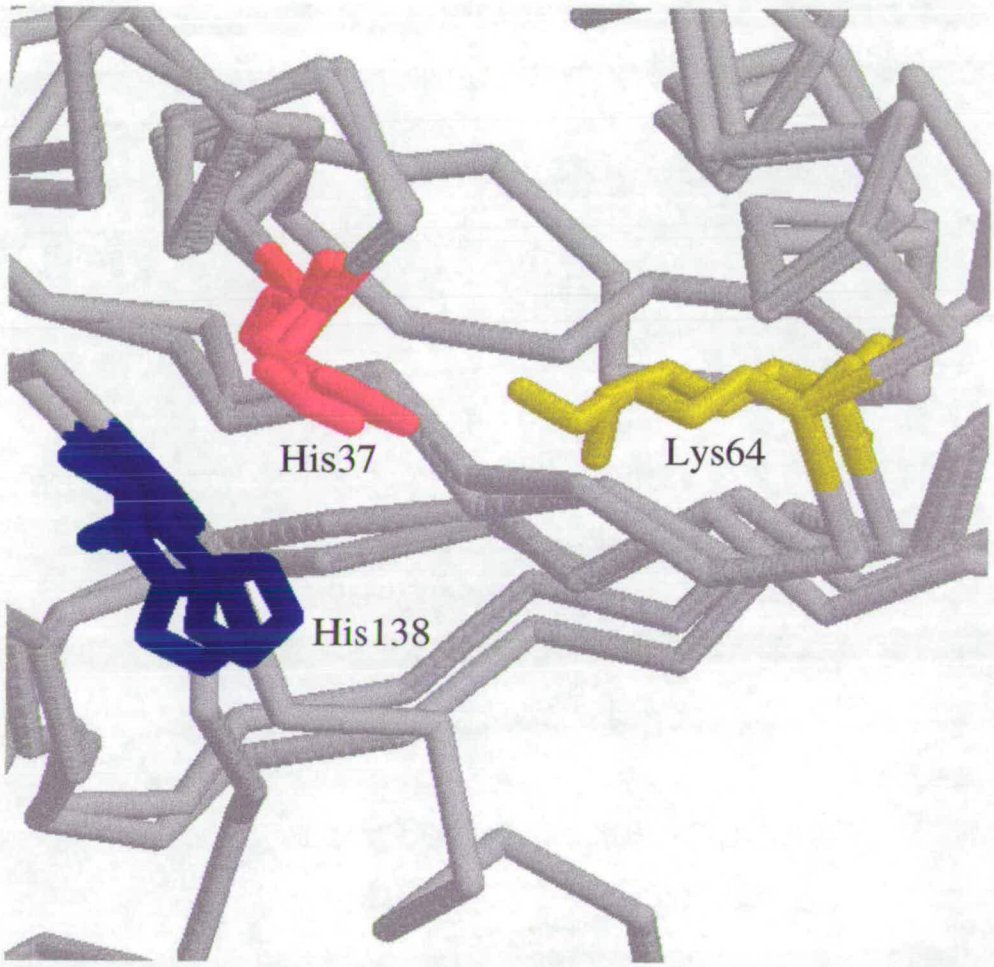


Figure 1.9: *Close up of the active site of human angiogenin showing similarity in predicted and actual residues.*

It is possible to parallelise the double dynamic algorithm which has shown an ability to detect similarities in structure beyond what is currently possible with sequence homology methods. Nevertheless, homology searching with the standard dynamic algorithm will always be far more efficient for database searching and means of improving the sensitivity of such searches need careful investigation.

The most important factor though appears to be the scoring table used. While global scoring schemes such as Dayhoff PAMs and Henikoff BLOSUM tables are suited to searches with no known structure involved, it is possible to extend the positional features of scoring by assessing each individual position in a sequence. Naturally, searches of the plain sequence database with a structure using a scoring based on the possible replacements in that structure would be useful and easily expressed as a profile. However, a more useful approach would be to have a full database of structures to search sequence against. Each individual entry would have a table tailored to its structure. Such a scoring scheme is described in Chapter 4.

Chapter 2

Automatic modelling

2.1 Introduction

The structure of a protein fold is directly determined by its primary sequence and as such it should be possible to compute the fold for any protein sequence. However, this is such a computationally intensive problem that it may never be solved using brute force calculation.

As seen in section 1.6 it is possible to build a respectable model of a new protein based on an alignment against a known structure. The aim of this chapter is to discuss the building of such models and how successful attempts at automatic modelling can expect to be.

2.2 Structure mutation

It is possible to use a known structure as the basis for a model of the new protein by preserving the backbone and unchanged residues and 'mutating' the residues where differences between the sequences exist. A sequence alignment provides a map of how one sequence may be changed into another (Figure 2.1). This mapping involves replacing residues and the introduction of gaps if necessary.

In this case the two sequences are very similar and so this should be an easy modelling case. There are no gaps to deal with, just changing five of the residues. A comparison of the two structures can be seen in Figure 2.2. Despite the very

```
           ********** ********.*.*..********
Db     121 GDFGADAQGAMTKALELFRNDIAAKYKELGFQG 153
Qy     121 GDFGADAQGAMNKALELFRKDMASNYKELGFQG 153
```

Figure 2.1: *Alignment of two short sections of similar polypeptide chain from horse myoglobin* (`Db`) *and human myoglobin* (`Qy`).

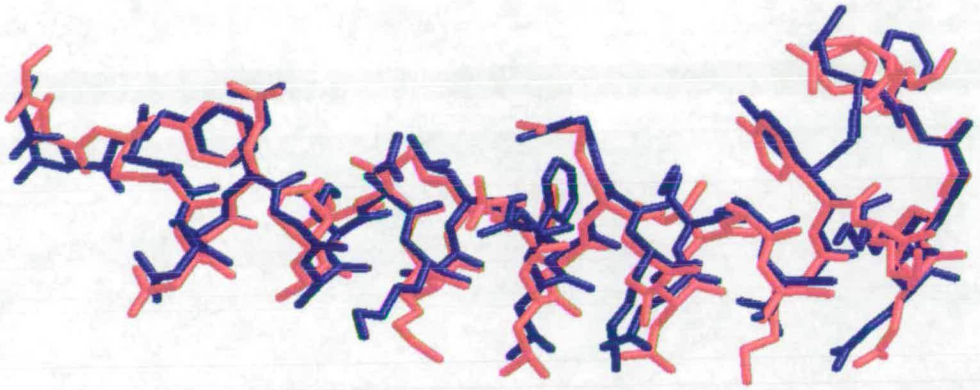


Figure 2.2: *Alignment of the structures of the two sequences in Figure 2.1 with Horse myoglobin in red and Human myoglobin in blue.*

high identity there are obvious differences in the conserved sidechains and the backbone which would have an impact on the ability to perfectly model one sequence from the other and these differences will have a cumulative effect on the model. Nevertheless, any model built by mutating one of these sequences into another is liable to be largely correct for the backbone and have a reasonable chance for the sidechains to be close to their correct position too.

An experiment to automatically mutate one helix into another was attempted using a simple set of sidechains and allowing rotation about the $C\alpha - C\beta$ bond to prevent clashes of atoms. In order to do this a program was written (`mutate`) which could load a Brookhaven PDB file into memory and replace the sidechains based on an alignment like Figure 2.1.

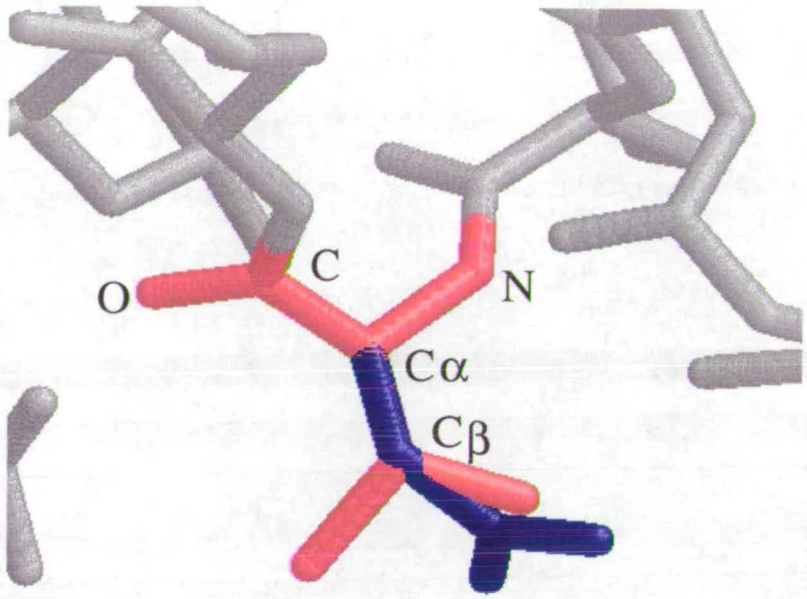


Figure 2.3: *Automatic replacement of thr (red) with asn (blue) showing how the orientation of the new sidechain is fixed according to the* $N - C\alpha - C = 0$ *of the original sidechain.*

2.3 Program development

A bottom-up approach for the code development was decided upon in which simple three-dimensional matrix routines for handling coordinate manipulations were built up into more and more complex subroutines. A single call is then sufficient to substitute one residue for another. The stages involved in this process are:

- Identify the backbone position of the old and new residues.
- Bring the backbones into line by translating the atoms of the new residue so that the $C\alpha$ atoms coincide.
- Rotate about the $C\alpha$ atom so that the backbone N atoms coincide.
- Rotate about the $N - C\alpha$ bond to make the carbonyl groups coincide.

As a result of this the new sidechain sits correctly relative to the backbone atoms but most have a range of potential conformations. Figure 2.3 shows the replacement of thr with asn.

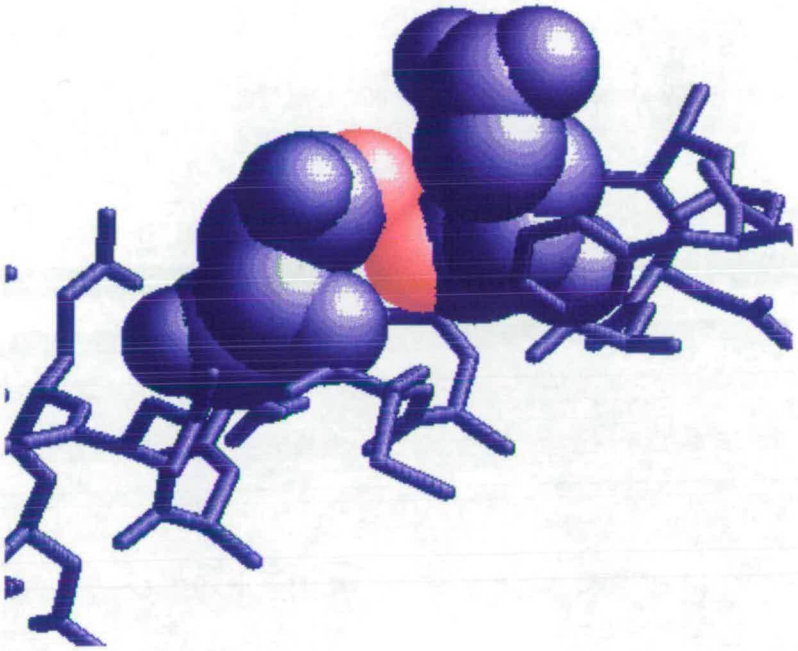


Figure 2.4: *Example of clashing atoms where the two red atoms overlap in space even though they are not covalently bonded.*

Examination of the atoms in the structure after a simple mutation reveals that many new sidechains are incorrectly placed resulting in their coordinates clashing with those of other atoms. To identify these atoms subroutines were written which examined all atoms and checked if their Van der Waals radii overlapped. Figure 2.4 shows the mutated thr which is now asn. Unfortunately, the nitrogen of the sidechain is overlapping with a carboxyl oxygen. The simplest fix is to rotate about the $C\alpha - C\beta$ bond. This relieves the clash as can be seen in Figure 2.5.

A comparison of this sidechain in both the model and real structure for human myoglobin can be seen in Figure 2.6. This replacement can be considered a success because it is very close to the correct orientation such that it is within the level of variation between conserved sidechains from the two structures.

Although the position of the sidechain is now quite acceptable this approach will not work for all sidechains, particularly those which are longer than this example, eg. Arginine. To correct any clashes occurring with such a sidechain it would be necessary to consider each of the rotatable bonds.

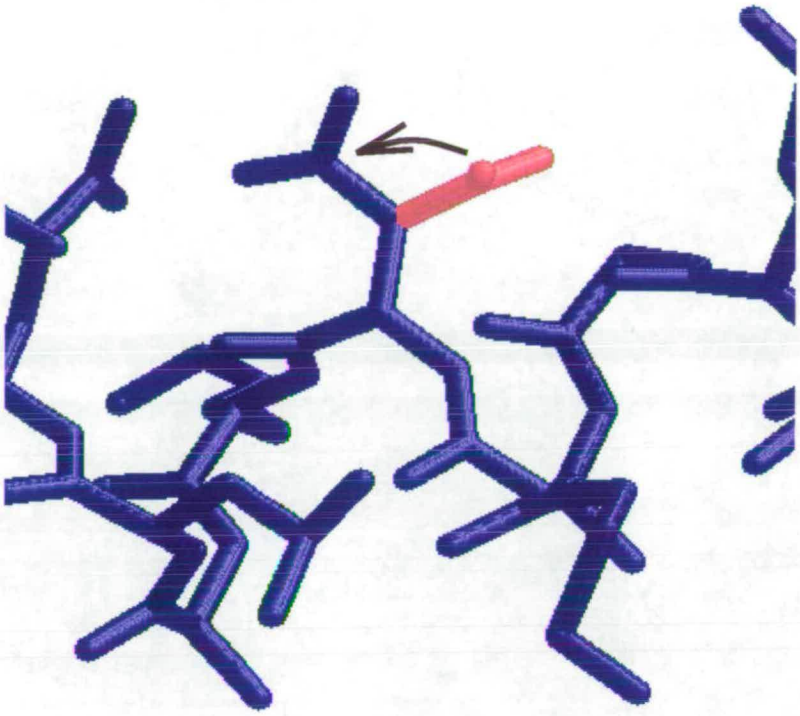


Figure 2.5: *The sidechain with the clash has been automatically rotated about the $C\alpha - C\beta$ bond. The unrotated sidechain is shown in red and the rotated version in blue.*

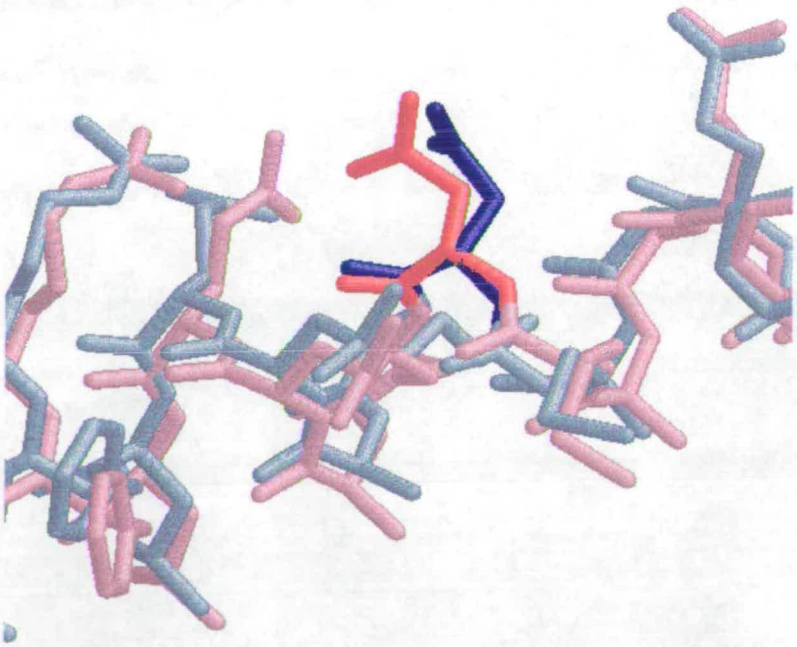


Figure 2.6: *A comparison of the model (red) for asn and the real (blue) orientation of the sidechain from Human myoglobin.*

In principle it would be possible to try rotating every sidechain about each of the rotatable bonds in the sidechains and at each step examine the clashes, but in practical cases the number of combinations is too large to be examined in a reasonable time (combinatorial explosion).

In order to reduce this problem another program has been used to search coordinate files for examples of real sidechains to build up a library of unique examples of every sidechain type. This rotamer library can then be used by the mutating program (Holm & Sanders, 1991).

In collecting the data a random set of PDB files was used and a program written using the various subroutines already built to search this set for examples of each type of residue. Initially a set of extended examples for each type was defined and as the program ran it compared each new example found with those already seen. The new example was aligned with each of those already known (initially just the extended form) and a comparison of the positions of the atoms made. If any of them had a root mean deviation of less than 1.5 Å they were considered identical to the new example and the number of that type seen incremented by one. Any new examples were then added to the database of known examples. This process continued until all examples of each residue in the database were compared.

Figure 2.7 is an example rotamer set generated for Cysteine. The total number of examples examined was 336 of which only 3 examples were stored. The first example is by far the most common seen and similar results were obtained with all residue types. The rotamer library frequencies for each sidechain are comparable to the results found by Ponder & Richards (1987). In each case, the first example was the extended form and this occurs most frequently.

Contrast this small set where the only freedom of movement is about the $C\alpha - C\beta$ bond with that of arginine which has a total of five rotatable bonds (Figure 2.8).

A good test for the rotamer library is to take a known structure and replace all of the sidechains with the most commonly found rotamers. This is what has been done (Figure 2.9) with the structure of ribonuclease. The first image shows the

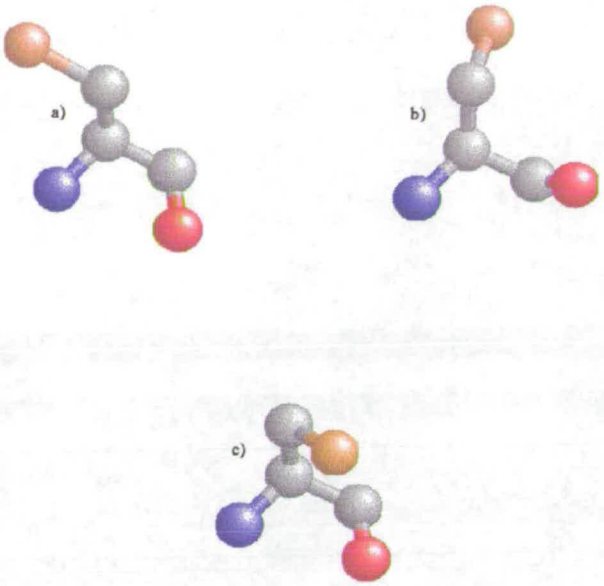


Figure 2.7: *Example rotamer set for cysteine showing the three examples found.*

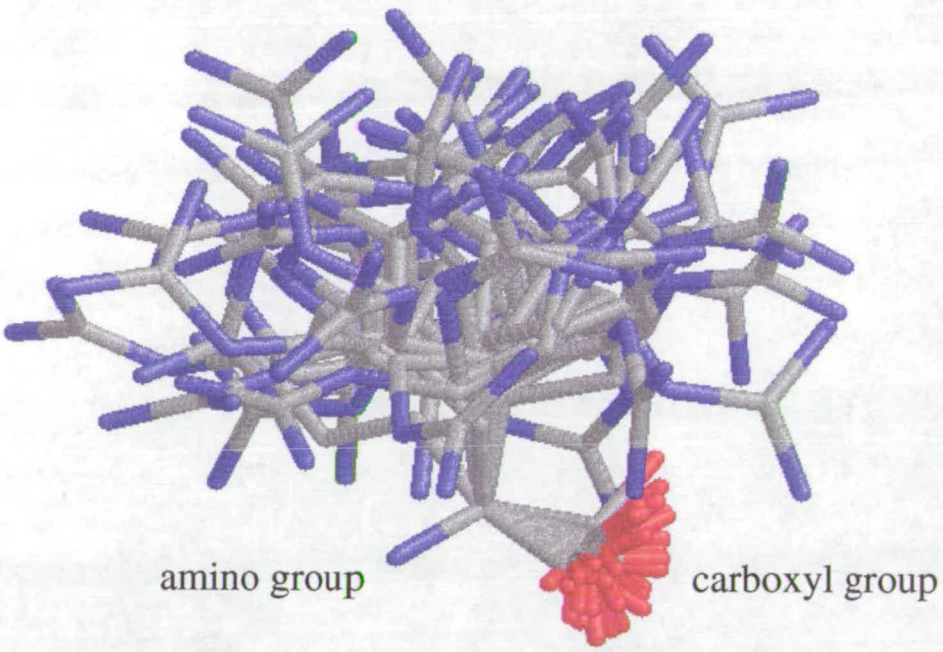


Figure 2.8: *Example rotamer set for arginine containing 49 examples.*

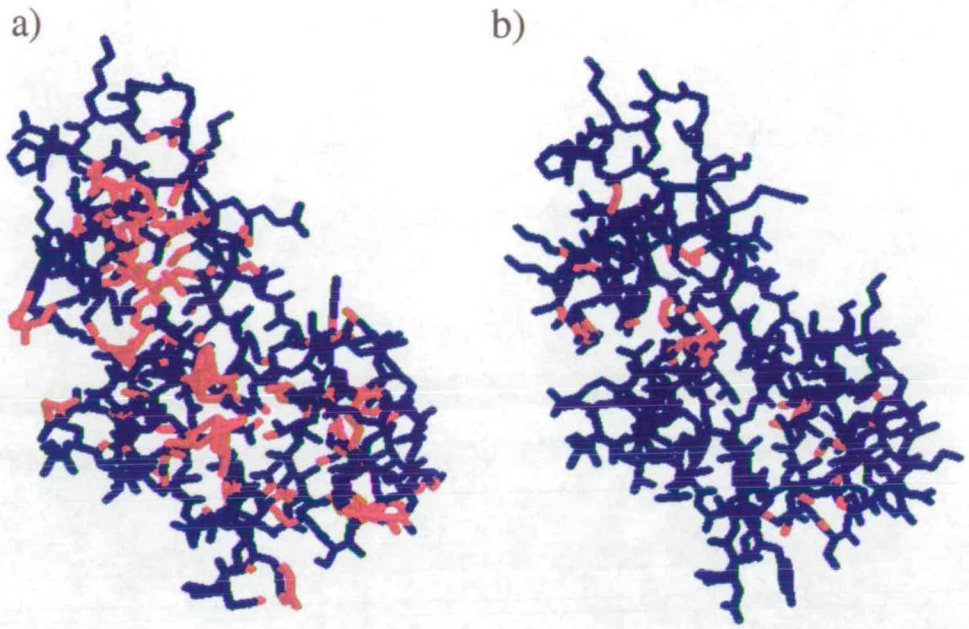


Figure 2.9: *Example of automatically mutated ribonuclease structure before* (a) *and after* (b) *the use of the rotamer library.*

structure where all sidechains are the extended versions, the second was produced using the rotamer library. The first has a lot of clashing atoms and the second shows far less but the process has not corrected all clashes.

2.4 Discussion

All that is required for the program to work is a template structure and a map of modifications in the form of a sequence alignment. The output is in the form of a Brookhaven PDB file and can be viewed using a molecular rendering program.

The use of a rotamer library improves the performance in reducing Van der Waals clashes. The present version of the program performs the mutations sequentially, from the N terminus to C terminus. This is not the ideal way to do it since the choice of rotamer at each position affects those chosen later. It may be possible to refine the replacement process by methods such as iterating through all possible combinations of available rotamers and picking the model with the smallest number of clashes. It is unlikely that this would result in a correct struc-

ture just based on clashing Van der Waals radii and may result in a combinatorial explosion as the number of sidechains increases. This would also depend on the size of the rotamer sets for each sidechain.

Dunbrack & Karplus (1993) extended the size of their rotamer libraries by taking account of backbone ϕ and ψ angles and this could improve the performance of `mutate` if implemented.

Eisenmenger *et al.* (1993) showed that as the number of identities between a known structure and the homologue to be modelled decreases, the accuracy of the prediction drops rapidly such that an all atom model based on an alignment with 60% identity is about 30% accurate. The probable reason is that if the backbone is wrong the sidechain replaced is likely to be wrong too. In addition, this assumes that the alignment is correct and if that is not the case then any structure based on that alignment has no chance of being correct.

In addition, the problem of insertions and deletions has not been considered in this chapter. Efforts to predict loops based on loop libraries have met with some success (vanVlijmen & Karplus, 1997). The effect of an inaccurate backbone model on the success of sidechain prediction makes the prediction of an all atom model structure unlikely. For now the way to get an all atom model is by crystal or NMR work.

While all atom models are of limited reliability, the ability to locate residues within elements of secondary structure does identify which residues may be important within a primary sequence and can guide experimental work. An example of this is described in the next chapter.

Chapter 3

Structural modelling of a type I DNA methyltransferase

3.1 Introduction

The function of a bacterial DNA restriction/modification (R-M) system is to maintain the methylation state of the chromosome by methylating the hemimethylated DNA produced during replication. In addition it restricts viral propagation by cleaving unmodified viral DNA. In a type I R-M system, the restriction endonuclease, modification methylase and sequence recognition functions require different subunits (R, M and S) in one large multifunctional enzyme (Wilson & Murray, 1991, Barcus & Murray, 1995). Type I R-M systems of enteric bacteria have been grouped into four families based on subunit complementation, DNA hybridisation and antibody cross-reactivity experiments (Barcus & Murray, 1995, Titheradge *et al.* 1996). The amino acid sequence identity is very high within a family for the R, M and the S subunit conserved regions (Gann *et al.* 1987, Cowan *et al.* 1989, Kannan *et al.* 1989, Sharp *et al.* 1992, Murray *et al.* 1993, Gubler et al. 1992). The variable regions of the S-subunit are the DNA target recognition domains (TRDs).

The type I S-subunit contains two TRDs made up of approximately 150–180 amino acids which each recognise one part of the bipartite targets (Bickle & Kruger, 1993, King & Murray, 1994, Barcus & Murray, 1995). Sequence conservation between TRDs is either below 20% for TRDs recognising different targets, or 40% (TRDs from a different family) to 90%+ (same family) when a target is shared. The conserved regions are responsible for defining the length of the

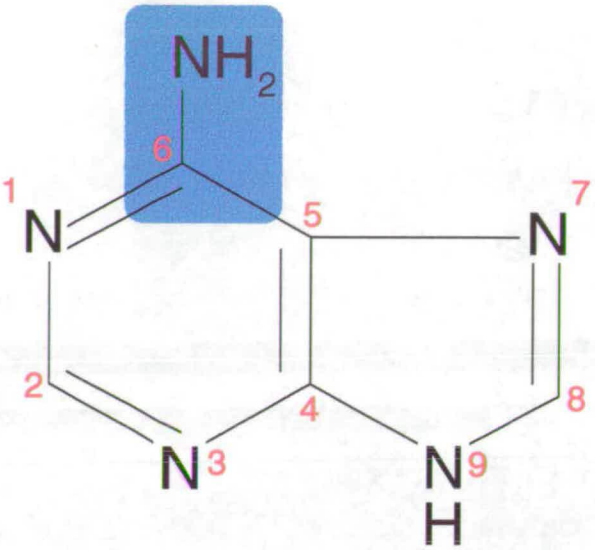


Figure 3.1: *Structure of adenine highlighting the N6 position.*

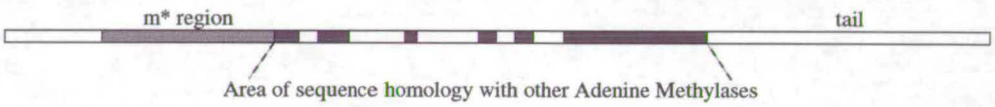


Figure 3.2: *Region of M-subunit modelled.*

non-specific DNA spacer in between the two TRD target sequences (Price *et al.* 1989) and for binding the M and R-subunits (Abadjieva *et al.* 1993, Meister *et al.* 1993, Cooper & Dryden, 1994, Webb *et al.* 1995). The TRDs of type I S-subunits recognise a wide variety of 3, 4 or 5 base pair targets and it is of interest to define which amino acids within the large and highly variable sequence of the TRDs are responsible for sequence specificity.

The M-subunit methylates the adenine nucleotide at the N6 position (Figure 3.1), one on each strand, using S-adenosyl methionine (SAM) as cofactor. In Eco$K$ it contains an N-terminal m* region and a C-terminal tail region which together give the preference for methylating hemimethylated targets rather than unmodified ones (Kelleher *et al.* 1991, Cooper & Dryden, 1994) (Figure 3.2)

The R-subunit cleaves DNA at an apparently random site remote from the target sequence after extensive DNA translocation driven by ATP hydrolysis (Wilson & Murray, 1991, Barcus & Murray, 1995).

```
Motif                 1'                  1                   2           3           4
M.EcoKI       (146) GQYFTPRPLIKTI (11) VQDPAAGTAGFLIEA (26) FIGLELV (25) IRLGNTL (08) KAHIVATNPPF (17)
M.EcoAI       (161) GEFYTPRAVTRFM (11) IMDPACGTGGFLACA (21) IHGVEKK (21) IRHDNTL (09) QLDVIVTNPPF (19)
M.EcoR124I    (196) GEFFTPQHVSKLI (13) IYDPAAGSGSLLLQA (12) FFGQEIN (22) IKLGNTL (09) PFDAIVSNPPY (28)
M.Mycoplasma  (195) GEFYTPSKVSELL (13) AYDPACGSGSLLIKL (09) IYGQEVK (22) LRSGDTL (10) SFDCIVANPPF (26)
M.TaqI        (018) GRVETPPEVVDFM (11) VLEPACADGPFLRAF (09) FVGVEID (11) GILADFL (06) AFDLILGNPPY (32)

Motif                                                 5
M.EcoKI       (17) NKQLCFMQHIIETL HPGGRAAVVVPDNVLF  EGGKGTDIRRDLMDKCHLHTILRLPTGIFYA QGVKTNVLFFTK (165)
M.EcoAI       (19) ETADLFLQLIVEVL AKNGRAAVVLPDGTLF  GEGVKTKIKKLLTEECNLHTIVRLPNGVFNPYTGIKTNLLFFTK (120)
M.EcoR124I    (28) KADFAFVLHALNYL SAKGRAAIVCFPGIFY  RGGAEQKIRQYLVDNNYVETVISLAPNLFFG TTIAVNILVLSK (114)
M.Mycoplasma  (26) YADFAFLQHMLFHVNKDNGIIASVFSLGILSRKSPKAEEDIRKYIIDKNYIDTIIFLPPNLFYN TSIESCIIVARK (116)
M.TaqI        (32) NLYGAFLEKAVRLL KPGGVLVFVVPATWLV   LEDFALLREFLARE   KTSVYYLGEVFPQ KKVSAVVIRFQK (211)
```

Figure 3.3: *Sequence alignment showing six conserved blocks.*

The subunit stochiometry is $R_2M_2S_1$ based on *in vitro* studies (Dryden *et al.* 1997).

3.2 Modelling the M-subunit of *Eco*KI

A comparison of all then known type I M-subunit sequences was made against the sequences of the γ class of type II N6 adenine methylases (Malone *et al.* 1995) (Figure 3.3). This identified the same six sequence motifs as found in the type II γ class methylase (Noyer-Weidner *et al.* 1994, Malone *et al.* 1995). Cheng (1995) showed that the catalytic domain of HhaI and *Taq*I exhibited a very similar three-dimensional folding, and suggested that this may follow for many SAM dependent methyltransferase catalytic domains.

Secondary structure predictions over the 200 amino-acid region containing the six conserved blocks show α-helix/β-strand/α-helix repeats over the whole region (Figure 3.4). The structural features predicted by the PredictProtein `PHD` program (Rost & Sander, 1993) were aligned by hand (Dryden *et al.* 1995) with those actually found in the crystal structure of the *Taq*I methylase catalytic domain (Labahn *et al.* 1994). Insertions and deletions in the type I sequences were found to be in surface loops in the *Taq*I structure. This alignment suggested the catalytic domain of type I systems was virtually identical to that of *Taq*I. Based on this alignment a model of the type I M-subunit of *Eco*KI was produced using Tripos `sybyl` (Figure 3.5).

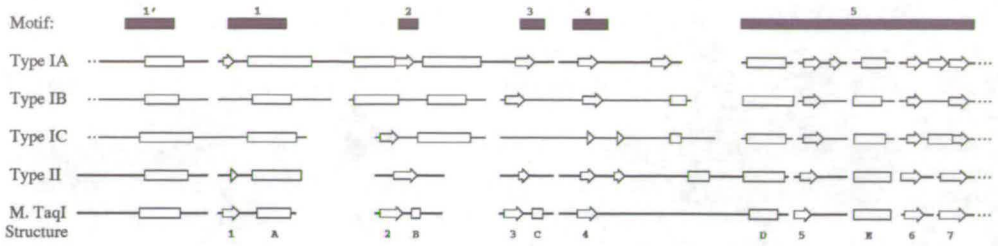


Figure 3.4: *Secondary structure predictions.*

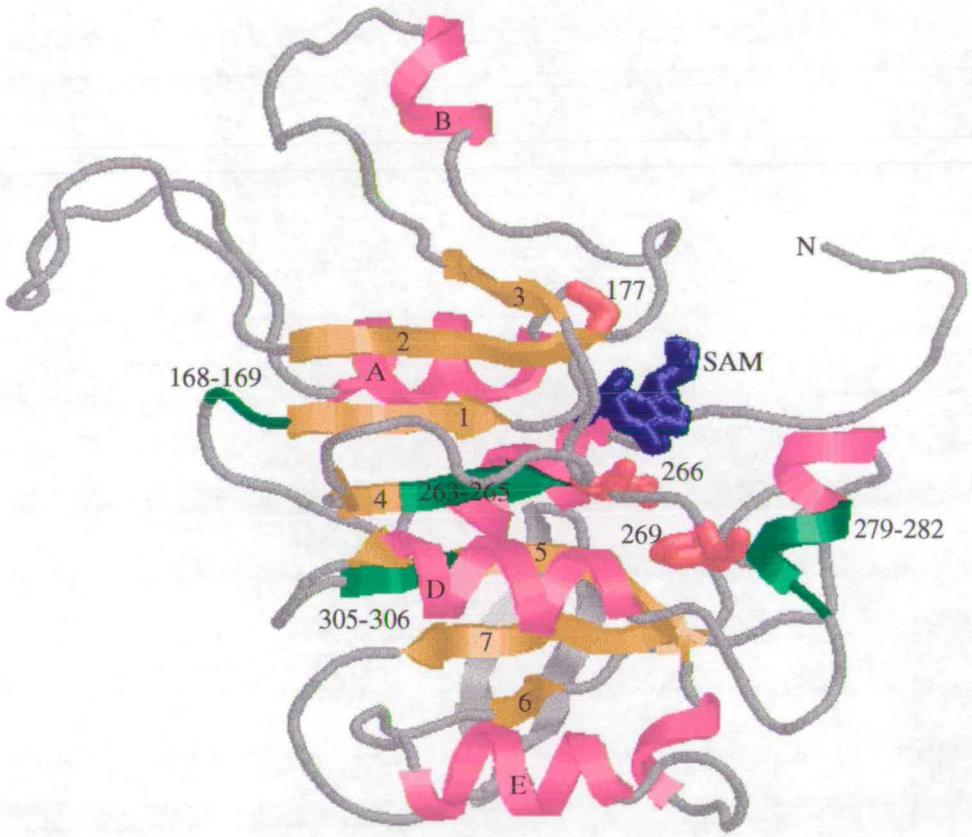


Figure 3.5: *A tertiary structure model of the catalytic domain from amino acids 141 to 380 in the M-subunit of Eco*KI *based on the sequence alignment in Figure 3.3 and the structure of the Taq*I *methylase. The α-helices, β-strands and SAM are coloured magenta, yellow and blue respectively. The helices and strands are numbered as in Figure 3.4. The sites of mutagenesis and proteolysis are coloured red and green respectively and numbered as in Table 3.1.*

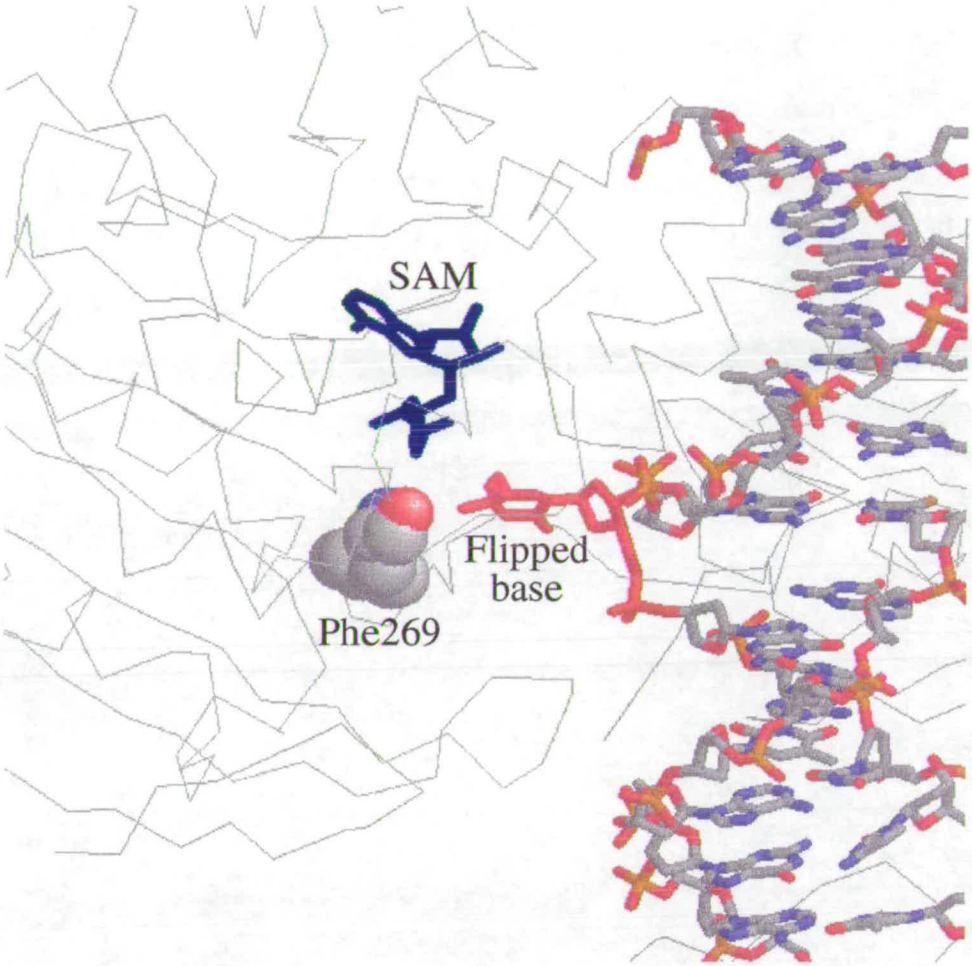


Figure 3.6: *Close up showing flipped out base in EcoK model.* Phe269 *and cofactor* (SAM) *are found here.*

3.2.1 Discussion

The model domain indicates plausible locations for the sites of mutagenesis (Willcock *et al.* 1994) and proteolysis with elastase, trypsin and chymotrypsin (Cooper & Dryden, 1994). In *Eco*KI (Table 3.1), Gly177 (motif 1) and Asn266 (motif 4) are located close to the cofactor and phe269 (motif 4) is found at the edge of the active site where it can interact with the target base if it is flipped out (Figure 3.6) of the DNA helix. Base flipping is thought to be quite widespread with strong evidence for it in Uracil-DNA glycosylase and *E.coli* photolyase with more examples expected to appear (Roberts, 1995).

Proteolysis yields stable polypeptides from the cleavage sites to the C-terminus

of the M-subunit (Cooper & Dryden, 1994). The model comprised two independant subassemblies joined by a loop between β-strands 3 and 4. The first subassembly has the form of the Rossman mononucleotide binding fold (Rossman *et al.* 1975). The second subassembly is similar (Malone *et al.* 1995) except for antiparallel β-strands 6 and 7, near the end of the domain. Proteases can only reach the buried sites on the β-sheet when the *N*-terminal half of the subunit containing the Rossman fold has been digested and the SAM binding site lost. The exposed loop containing the Arg279–Val282 site is protected by SAM binding from digestion by trypsin and chymotrypsin, suggesting a conformational change possibly analogous to that observed for the similar loop in *Hha*I methylase (Klimasauskas *et al.* 1994).

Based on the suggestions that a type I methylase has approximate two fold rotational symmetry (Cooper & Dryden, 1994, Meister *et al.* 1993, Taylor *et al.* 1994), a model of the entire methylase part of a type I system was made bound to DNA using two copies of the type II *Hha*I methylase structure (Figure 3.7). The catalytic domain of *Hha*I was substituted with the model for this figure. The remainder of the structure is unaltered *Hha*I and is only meant to indicate the probably position and shape of the S subunit.

The location of the start of the modelled domain, which contains two m* mutation sites at ser144 and arg153 (Kelleher *et al.* 1991) suggests that the *N*-terminal m* region is folded up against the back of the catalytic domain. The tail of the M-subunit is sensitive to the methylation state of the DNA target (Cooper & Dryden, 1994). To allow one catalytic domain, bound at one half of the DNA target, to be sensitive to the methylation state of the other half of the target (type IA methyltransferases prefer hemi-methylated DNA), it appears necessary to place the tail of each M-subunit against the DNA and TRD forming the second half of the target site (Figure 3.8). In the absence of further structural information, the m* region and the tail region are shown as a single block for each M-subunit (Dryden *et al.* 1995).

The model is consistent with biochemical results, including DNA footprinting (Powell & Murray, 1995) and protein-DNA crosslinking (Chen *et al.* 1995) experi-

Proteolytic cleavage sites in catalytic domain	**Comment** (Cooper & Dryden, 1994)
R168 ↓ E169	Trypsin cleavage at partially exposed loop, fragment from amino acid 169 to ≈ 440. On β-strand 1.
V263 ↓ A264 A264 ↓ T265 R305 ↓ A306	Elastase or trypsin cleavage at buried site on β-strand. Fragments from cleavage site to end of M-subunit. On β-strands 4 and 5.
R279 ↓ T280 F281 ↓ V282	Trypsin or chymotrypsin cleavage at exposed loop. Fragments from cleavage site to end of M-subunit. SAM binding prevents cleavage and stabilises whole subunit. On loop between β-strand 4 and α-helix D.
Site-directed mutations	(Willcock *et al.* 1994)
G177D	Abolishes SAM binding, insoluble when expressed at 37^{o}C, soluble at 25^{o}C. Methylase structure and DNA binding normal. In motif 1 between β-strand 1 and α-helix A.
N266D, F269G, F269C	SAM and DNA binding normal but inactive. In motif 4 at end of β-strand 4.
F269Y, F269W	SAM and DNA binding normal. Activity 25% and $\leq$5% of wild type mtase respectively.

Table 3.1: *Sites of proteolysis and mutation in the putative catalytic domain of Eco*KI.

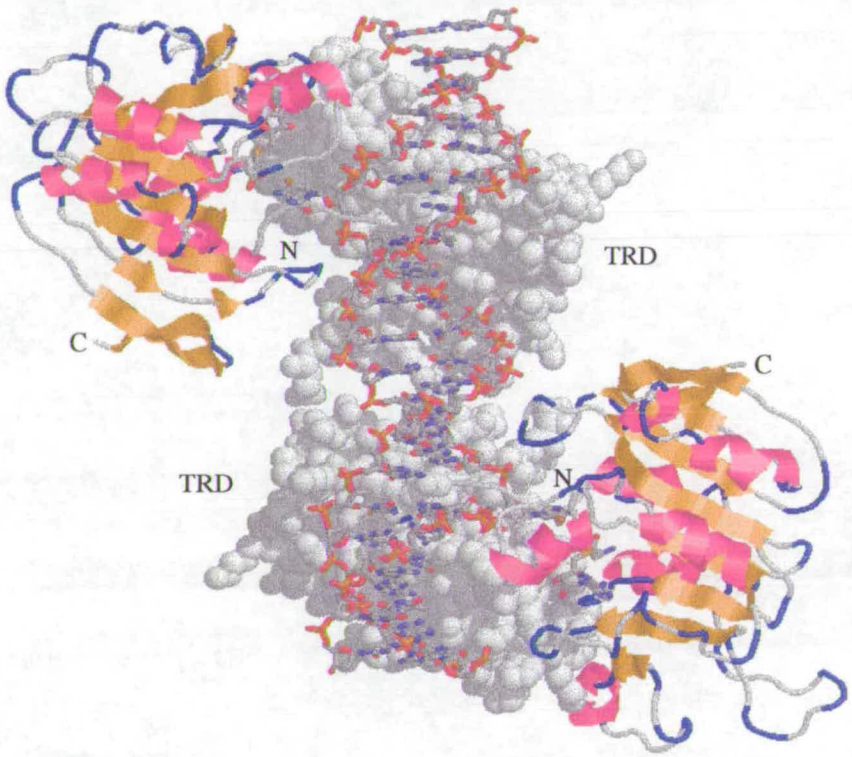


Figure 3.7: *A front view of a partial model of a type I methylase bound to DNA. The target bases seen flipped out are 10 bases apart as found in the Eco*KI *recognition sequence.*

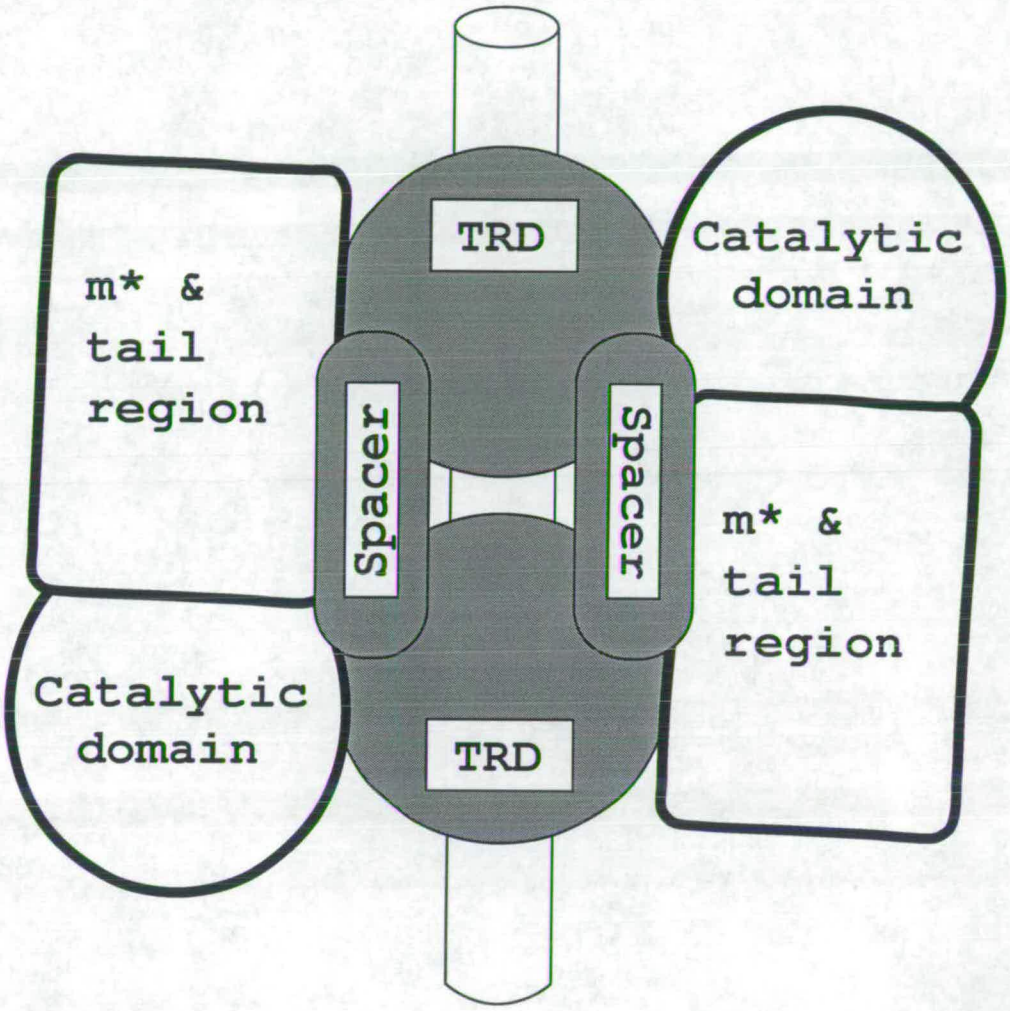


Figure 3.8: *A schematic rear view of the complete methylase structure bound to DNA. The M-subunits are shown in a bold outline with the postulated position of the N-terminal m* and C-terminal tail regions and the catalytic domain indicated. The S-subunit is shown shaded with the TRD joined by the two short spacer regions which are well conserved in all S-subunit families. The DNA helix, shown as a cylinder, lies on the S-subunit and between the M-subunits in agreement with earlier models (Taylor et al. 1994, Burckhardt et al. 1981).*

ments, and with measurements of the hydrodynamic shape of the protein (Taylor *et al.* 1994, Powell *et al.* 1993).

3.3 S-subunit TRD alignment and modelling

For this work most nucleotide or amino acid sequences of the S-subunits were obtained from published references or GenBank except for the S-subunits of *Bsu*CI and *Kpn*AI which were provided by Prof. T. Trautner (Berlin) and Dr. J. Ryu (Loma Linda).

Previous attempts to align S-subunits have led to unconvincing results (Argos, 1985, Gann *et al.* 1987) principally because of the few sequences available at the time, and also because of the high degree of sequence similarity exhibited by the conserved regions. This meant that very little could be said about the alignment of the TRD regions except that they were highly variable. In order to get over this problem the putative TRDs were collected into a database excluding the conserved regions. The locations of the TRDs in the S-subunit sequence and, if known, their DNA target and type I families are given in Table 3.2.

The database was searched with each member (inverted) using `sss_align` with sequence data only. This gave a set of close homologues to each sequence which was used as the input to the PredictProtein `PHD` server. Many of these sequences were not in SwissProt so `PHD` could not find them. Some of the sequences did not appear to have any relatives with greater than 30% identity and these sequences were submitted alone. Predictions with these sequences have a much lower expected accuracy for the secondary structure prediction.

The results from `PHD` were then placed in another database and this was inverted using `sss_align`. The addition of secondary structure information meant that it was possible to align more distant relatives. The results were used to cluster sequences and overlap previous clusters until the alignment (Figure 3.9) included nearly all of the TRD sequences. Some TRD sequences could not be inserted into this alignment by `sss_align` because their secondary structure predictions were

Family	Name*	DNA target	Length	1st TRD	2nd TRD
IA	*Eco*KI	AAC N6 GTGC	464	11–157	214–368
IA	*Eco*BI	TGA N8 TGCT	474	11–158	215–379
IA	*Eco*DI	TTA N7 GTCY	444	11–128	185–348
IA	*Sty*LTIII	GAG N6 RTAYG	469	11–153	209–375
IA	*Sty*SPI	AAC N6 GTRC	463	11–157	214–367
IA	*Eco*R5I		>140	1–140	
IA	*Eco*R10I		>131	1–131	
IA	*Eco*R13I		>152	1–152	
IB	*Eco*AI	GAG N7 GTCA	589	110–247	403–540
IB	*Eco*EI	GAG N7 ATGC	594	109–247	403–545
IB	*Cfr*AI	GCA N8 GTGG	578	108–236	385–529
IB	*Sty*SKI	CGAT N7 GTTA	587	108–249	396–538
IB	*Sty*STI		>146	1–146	
IB	*Eco*R17I	ATR.....	>138	1–138	
IC	*Eco*R124I	GAA N6 RTCG	409	24–142	215–350
IC	*Eco*DXXI	TCA N7 RTTC	406	23–139	211–341
IC	*Eco*prrI	CCA N7 RTGC	405	22–159	232–360
ID	*Sty*SBLI	CGA N6 TACC	434	1–153	229–405
ID	*Eco*R9I		464	1–188	264–435
IC?	*Bsu*CI	GAY N7 TGGA	405	23–162	219–355
IC?	*Kpn*AI		439	1–155	231–410
IC?	*Mpu*1AI		401	1–139	221–359
IC?	*Mpu*1BI		336	1–139	188–324
ID?	HI0216		385	20–138	198–333
ID?	HI1286		459	1–176	268–445
?	mj0130		425	23–161	231–371
?	mj1218		425	28–155	226–368
?	mj1531		425	28–170	241–371

Table 3.2: *S-subunits of type I restriction/modification systems. (* see Table 3.3).*

Abbreviation	Full name
Eco	*Escherichia coli*
Sty	*Salmonella enterica*
Cfr	*Citrobacter freundii*
Bsu	*Bacillus subtilis*
Kpn	*Klebsiella pneumomiae*
Mpu	*Mycoplasma pulmonis*
HI	*Haemophilus influenzae*
mj	*Methanococcus jannaschi* gene number
Ngo	*Neisseria gonnorhoeae*

Table 3.3: *Table of abbreviations used in Table 3.2.*

Colour	Residue type
Red	Acidic
Blue	Basic
Green	Hydrophilic
Pink	Hydrophobic

Table 3.4: *Table of colours used in Figure 3.9.*

of low accuracy, the sequences were too highly diverged or a combination of the two. In these cases (indicated in Figure 3.9 by an asterisk after the TRD name) the sequences were inserted manually. The amino terminal TRD and carboxy terminal TRD are indicated by the suffix -1 or -2 appended to the system's name. Table 3.4 shows the key for the colour scheme used.

In addition to aligning the set of predictions against each other, they were all aligned against the known TRD crystal structure of *Hha*I and the best alignment was found against TRD *Eco*KI-1. This was then used as the key sequence to provide probable locations for the loops and strands involved in DNA recognition.

In contrast to previous analyses, sequence similarity between TRDs is visible even if the TRDs recognise different DNA targets. In the figure, sequences which are closely related to each other in the output of `sss_align` are close to each other

in the alignment and those further away are either very distant or unrecognisable as significant alignments. It is interesting to observe that the secondary structures for most of the sequences are very similar even though they were produced using independant groups of homologues.

In building the alignment, β-strand 1 was used as the centre point and was identified in all sequences. This was then used to lock the sequences together which resulted in the similarity of other secondary struture features being apparent, particularly loops 1 and 2 and β-strand 2.

3.3.1 Discussion

The agreement in length and composition of strand 1 is good between *Hha*I and all of the type I TRDs. Loop 1 is generally predicted to be shorter and β-strand 2 longer than the equivalent structures in *Hha*I. Figure 3.10 shows the recognition of the DNA phosphate backbone and bases by part of the TRD of *Hha*I methylase. In *Hha*I, loop 1 (val232–glu239) fills the major groove and positions gln237 into the gap left by the flipped out cytosine base, β-strand 2 (arg240–tyr242) makes important base and phosphate contacts, and thr250–phe259, as part of the long loop 2, makes further backbone and base contacts. Strand 2 commences with arg240 and it is apparent that an equivalent basic amino acid e.g. lys92 in *Eco*KI, is present in many of the type I TRDs. Arg240 in *Hha*I is involved in base recognition and perhaps suggests a similar role for these basic residues in type I S-subunits. Loop 2 is 21 amino acids long in the *Hha*I-DNA cocrystal structure but in the absence of DNA, the loop is interrupted by a β-strand at amino acids 250–253 (Cheng *et al.* 1993, Klimasauskas *et al.* 1994). The existence of an equivalent extra β-strand in the middle of loop 2 is predicted for many of the type I TRDs. In *Hha*I methylase, loop 2 terminates with another β-strand. However, many of our predictions suggest that in type I TRDs, loop 2 is followed by an α-helix while others suggest that it is a β-strand. This may suggest that the structure of type I TRDs deviates from that of *Hha*I at this junction. Conversely, if the structure of the rest of the model is correct there needs to be a β-strand at that point to

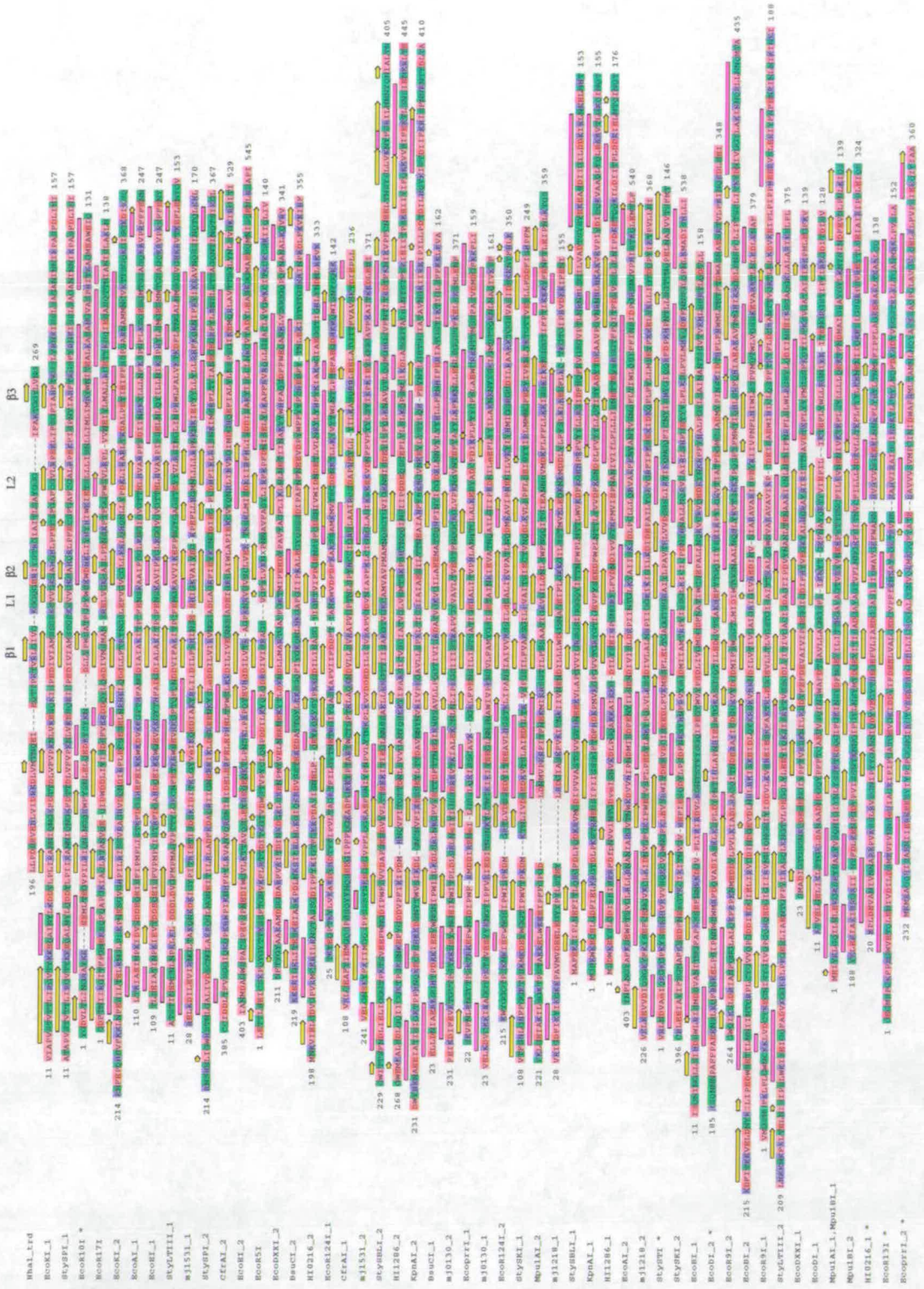


Figure 3.9: *Alignment of all available S subunit TRDs showing predicted secondary structures. The top entry, HhaI has a known structure and this region is shown in Figure 3.10.*

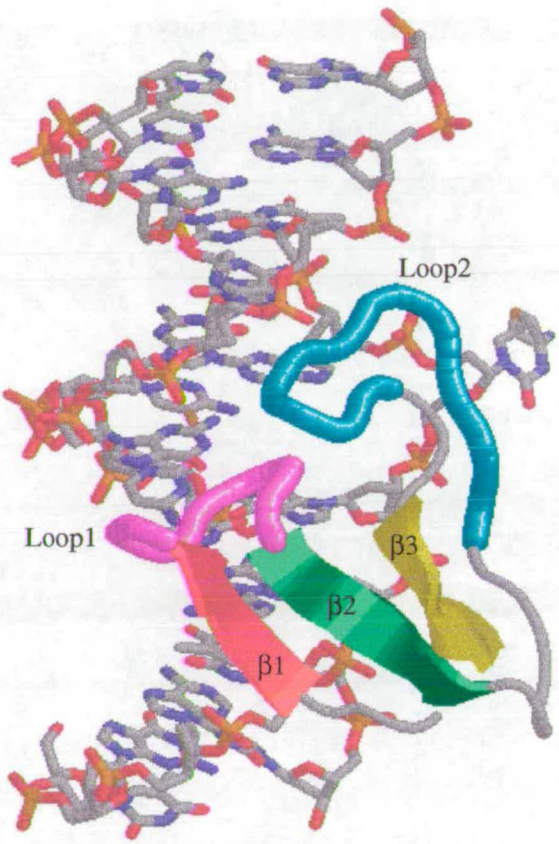


Figure 3.10: *Region of Hha*I *corresponding to the strands and loops seen in Figure 3.9.*

H-bond with β-strand 2 completing the sheet structure so it would be likely that the predictions for α-helix are incorrect at this point.

A variety of experiments such as UV-induced crosslinking to DNA (Chen *et al.* 1995), chemical modification of lysines (Taylor *et al.* 1996), and random mutagenesis of TRDs (personal communication, M O'Neill and N.E. Murray) have been applied to the best characterised type I R-M systems, *Eco*KI and *Eco*R124I, to identify amino acids involved in sequence recognition.

Chemical modification of *Eco*R124I showed that several lysines in the second TRD were susceptible to modification especially in the absence of bound DNA (Taylor *et al.* 1996). Lysines 261, 297 and 327 within the TRD were particularly strongly modified. Lys297 is the most strongly modified residue and lies within the second proposed recognition loop. These three lysine residues are also conserved in the first TRD of *Sty*SKI which recognises the same DNA target as the second TRD of *Eco*R124I therefore supporting a role for them in sequence recognition (Thorpe *et al.* 1997).

Random mutagenesis of the first TRD of *Eco*KI has changed 40 out of 150 amino acids. Most of the mutations are silent but three of five mutations that impair restriction and modification are within the two putative recognition loops.

UV crosslinking demonstrated that tyr27 in the first TRD of *Eco*K was in contact with the 3′ thymine base in the sequence complementary to the 5′ AAC part of the *Eco*K target (Chen *et al.* 1995). This residue is outside the predicted recognition loops. However, it has been found that changing it to other amino acids has a minor effect on DNA specificity suggesting that it may be involved in a non-sequence specific interaction with the DNA (personal communication, M. O'Neill and N.E. Murray).

3.4 Conclusion

This work strongly suggests an evolutionary link between all methylases in type I and type II R-M systems. Almost all of the known R-M systems are of the type II class which contain a simple modification methylase, but recent results suggest that the more complex type I R-M systems may also be widespread in nature (Barcus & Murray, 1995, Dybvig & Yu, 1994). The link between type I and II methylases may be extended to suggest that all of the DNA methylases are constructed from one or more copies of the catalytic domain and one or more TRDs. These domains can be on the same polypeptide chain or on separate subunits. In some DNA methylases — the C5 cytosine (Kumar *et al.* 1994), N6 adenine β-class (Wilson, 1992), and type III N6 adenine methylases (Wilson & Murray, 1991, Barcus & Murray, 1995) — the catalytic domain is interrupted by the TRD. This could be accommodated structurally due to the nature of the two substructures within the catalytic domain and could arise by gene duplication and rearrangement (Lauster, 1989, Malone *et al.* 1995). The presence of additional domains and subunits in type I and type III R-M systems compared to the type II R-M systems allows the acquisition of a more sophisticated range of enzymatic responses on binding to a DNA target.

The similarity between TRDs of type I N6-adenine methylases and the TRDs of C5-cytosine methylases may extend to many, if not all, TRDs of type II N6-adenine methylases, type III methylases and other methylases which do not fit current classifications. This would support and extend the proposal (Wilson & Murray, 1991) that all methylases have evolved from a common ancestor consisting of a small monomeric TRD, such as that still found in *Aqu*I methylase (Karreman & de Waard, 1990) and *Eco*HK31I methylase (Lee *et al.* 1995), associated with a separate catalytic subunit. It has been proposed that the methylase catalytic subunit may have developed from early DNA repair enzymes which use the same base flipping method to gain access to their target base as the methylases (Roberts, 1995). The phage-mediated selection with novel restriction-modification

systems favours rare genotypes to maintain genetic variability (Levin, 1988). This variability provides a bacterial population with the ability to invade bacteria-free habitats in which phage are present (Korona & Levin, 1993). The high mutation rate has virtually obscured the common origin for all TRDs. A conserved tertiary structure within TRDs implies that it may eventually be feasible to derive the amino acid recognition code used by TRDs to recognise DNA sequences.

Chapter 4

Sequence and secondary structure alignment

4.1 Introduction

The aim of homology modelling is to predict structures for new protein sequences based on alignment against sequences with known structures. This problem can be split into two areas, fold recognition and comparative modelling. This is not difficult with high homology sequences and the predicted structures for sequences which are 80% identical do have a good level of agreement with the true structure. But as identity drops the accuracy of alignments decreases.

Thus, the problems which need to be addressed before a model can be built are that the homologous sequence should be identified, and that the alignment should be accurate. The program `sss_align` approaches these two problems by using a combination of sequence and secondary structure information.

4.1.1 Secondary structure prediction alignment

Recent advances in the application of neural networks to protein secondary structure prediction (Rost & Sander, 1993) have resulted in predictions which regularly achieve over 70% accuracy. Therefore, it becomes possible to align sequences using

predicted secondary structure to try and detect similarities beyond the reach of sequence alignment and without resorting to the time consuming fold threading methods. TOPITS (Rost, 1995) and MAP (Russell *et al.* 1996) attempt to do this.

TOPITS projects a database of known three-dimensional structures onto one-dimensional strings of secondary structure and relative solvent accessibility. The `PHD` program is then used to produce similar strings for a sequence of unknown structure. This prediction is then aligned by dynamic programming against the database. Rost also used a combination of this method with sequence analysis to improve the performance in some cases.

The MAP program is very simple in that it simply converts the database entries and prediction into strings of letters which can be aligned by dynamic programming. After this crude database search, various heurisms are applied to discard hits which are impossible structurally. For example, aligning two secondary structure elements when there is not enough peptide chain between the two to bridge the gap. After this, the user is presented with a list of biologically plausible structures which fit the prediction.

4.1.2 Sequence and secondary structure alignment

Sequence alignment is still the primary means of comparing proteins. Given knowledge of secondary structure in addition to the sequence it should be possible to boost the score of sequence hits in a normal database search if the secondary structure of the database entry matches the prediction of the query, taking into account the likelihood of the prediction being wrong. This was the premise on which `sss_align` was built.

4.2 Program development

Previous work on dynamic programming using the Smith & Waterman best local similarity algorithm (`MPsrch`, Collins & Sturrock, unpublished) meant that the same algorithm was a good starting place to try out a new scoring scheme. It is well known how the Smith & Waterman algorithm behaves in database searches under many varied conditions so using this as a base line would make proving the new scoring scheme's effectiveness more simple.

4.2.1 Prediction reliability

Rost states that the reliability for the three state prediction of `PHD` is: 77.8% Helix; 64.5% Strand; 74.0% Loop. Given this information it is relatively easy to construct a log odds scoring scheme which will have values for matching like secondary structures and unlike secondary structures.

Substitute W, X, Y, Z in the following to be helix, strand, loop or turn for all combinations as appropriate. Calculate the average frequency of each type (substitute $\bar{A}$ for $\bar{W}, \bar{X}, \bar{Y}, \bar{Z}$) in the two sequences being matched (Equation 4.1). Then the values for *match* and *mismatch* are calculated using equations 4.2 and 4.3. While `PHD` does not actually provide a prediction for turns it was felt that it would be useful to have room for them for when the input sequences contain turns. If turns are not present in both sequences they are considered to be equivalent to loop residues where they do occur.

$$\bar{A} = \frac{fA_{seq1} + fA_{seq2}}{2} \tag{4.1}$$

$$match = \log_{10} \frac{\left(\begin{array}{c} \left(\bar{W} \cdot p\left(W_{seq1}\right) \cdot p\left(W_{seq2}\right)\right) \\ + \\ \left(\begin{array}{c} \bar{X} \cdot (1.0 - p\left(X_{seq1}\right)) \cdot (1.0 - p\left(X_{seq2}\right)) \cdot \\ \left(\frac{fW_{seq1}}{(fW_{seq1}+fY_{seq1}+fZ_{seq1})}\right) \cdot \\ \left(\frac{fW_{seq2}}{(fW_{seq2}+fY_{seq2}+fZ_{seq2})}\right) \end{array} \right) \\ + \\ \left(\begin{array}{c} \bar{Y} \cdot (1.0 - p\left(Y_{seq1}\right)) \cdot (1.0 - p\left(Y_{seq2}\right)) \cdot \\ \left(\frac{fW_{seq1}}{(fW_{seq1}+fX_{seq1}+fZ_{seq1})}\right) \cdot \\ \left(\frac{fW_{seq2}}{(fW_{seq2}+fX_{seq2}+fZ_{seq2})}\right) \end{array} \right) \\ + \\ \left(\begin{array}{c} \bar{Z} \cdot (1.0 - p\left(Z_{seq1}\right)) \cdot (1.0 - p\left(Z_{seq2}\right)) \cdot \\ \left(\frac{fW_{seq1}}{(fW_{seq1}+fX_{seq1}+fY_{seq1})}\right) \cdot \\ \left(\frac{fW_{seq2}}{(fW_{seq2}+fX_{seq2}+fY_{seq2})}\right) \end{array} \right) \end{array} \right)}{(fW_{seq1} \cdot fW_{seq2})} \tag{4.2}$$

$$mismatch = \log_{10} \frac{\left(\begin{array}{c} \left(\begin{array}{c} (fW_{seq1} \cdot p\left(W_{seq1}\right) \cdot (1.0 - p\left(X_{seq2}\right))) \cdot \\ \left(\frac{fX_{seq2}}{(fX_{seq2}+fY_{seq2}+fZ_{seq2})}\right) \end{array} \right) \\ + \\ \left(\begin{array}{c} (fX_{seq2} \cdot p\left(X_{seq2}\right) \cdot (1.0 - p\left(W_{seq1}\right))) \cdot \\ \left(\frac{fW_{seq1}}{(fW_{seq1}+fY_{seq1}+fZ_{seq1})}\right) \end{array} \right) \\ + \\ \left(\begin{array}{c} \bar{Y} \cdot (1.0 - p\left(Y_{seq1}\right)) \cdot (1.0 - p\left(Y_{seq2}\right)) \cdot \\ \left(\frac{fW_{seq1}}{(fW_{seq1}+fX_{seq1}+fZ_{seq1})}\right) \cdot \\ \left(\frac{fW_{seq2}}{(fW_{seq2}+fX_{seq2}+fZ_{seq2})}\right) \end{array} \right) \\ + \\ \left(\begin{array}{c} \bar{Z} \cdot (1.0 - p\left(Z_{seq1}\right)) \cdot (1.0 - p\left(Z_{seq2}\right)) \cdot \\ \left(\frac{fW_{seq1}}{(fW_{seq1}+fX_{seq1}+fY_{seq1})}\right) \cdot \\ \left(\frac{fW_{seq2}}{(fW_{seq2}+fX_{seq2}+fY_{seq2})}\right) \end{array} \right) \end{array} \right)}{(fW_{seq1} \cdot fX_{seq2})} \tag{4.3}$$

With these two equations it is possible to calculate a log odds score for two sequences with secondary structures, either predicted or real.

4.2.2 Merging sequence and secondary structure

Once the scoring scheme was built it was necessary to find a way to merge it seamlessly into a sequence search. Since normal Dayhoff PAM tables have a range of positive and negative scores which varies depending on the distance of the table selected, and the optimal gap penalty used is determined by the table, if a successful merging of scoring schemes was to be achieved, the secondary structure scoring scheme should have a range which corresponds to the sequence table so that the same gap penalty can be applied to each table. In practice this meant scaling the secondary structure table so that its most negative value matched the most negative value seen in the sequence table.

Once this scaling is achieved a simple way to merge the two is to take a percentage of one table and a percentage of the other and sum them. In the event this would leave a sliding scale going from 0% sequence and 100% secondary structure table to 100% sequence and 0% secondary structure. This allows the subtle inclusion of secondary structure information which will act as a very fine filter to capture significant hits bringing them out of the noise and drop insignificant hits into the noise.

The end result is a program which behaves like a normal dynamic Smith & Waterman algorithm retaining all the positive aspects of that method, while allowing the introduction of secondary structure information as needed to improve alignment quality and significance.

4.2.3 Fixed and variable scoring

Accepting the secondary structure reliability on a global fixed scale would limit the sensitivity of the alignment method. Although sequence homology would make up for deficiencies in the prediction to an extent, it would be preferable to exclude

```
          ....,....1....,....2....,....3....,....4....,....5....,....6
AA       |AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKV|
PHD sec  |  EEEEEEEEEEEEEEE  EEEEEE    EEEEEEEE           EEEE  EEEEEE|
Rel sec  |952665799999987313358998388732588986211201577321366285899999|

prH sec  |000000000000000000000000000000000013345441113331000000000000|
prE sec  |025776888989888553378998611135788887421000100002577512899999|
prL sec  |973222100000011345521001388753211001244344687653211487100000|
SUB sec  |LL.EEEEEEEEEEE....EEEEE.LLL..EEEEEE......LLL....EE.LLEEEEEE|
```

Figure 4.1: *Typical prediction output from* PHD.

Rel index:	0	1	2	3	4	5	6	7	8	9
Qhelix %pred	39.7	44.2	49.2	53.1	57.6	62.9	68.2	73.4	82.9	92.6
Qstrand %pred	39.8	39.5	45.0	50.8	55.2	60.0	66.8	73.4	81.1	89.6
Qloop %pred	51.5	51.5	54.6	60.1	65.1	68.8	74.3	78.8	85.5	91.4

Table 4.1: *Conversion table for* PHD *output to actual reliabilties.*

parts of the prediction which are known to be less reliable while boosting the interest in parts where accuracy is likely to be higher. Fortunately PHD provides suitable numbers for this in the form of the per residue reliabilities. These numbers are scaled from the actual probablilities and B. Rost provided the numbers to which they correspond (Figure 4.1 & Table 4.1).

Given these numbers for every position in the prediction they can be fed into the equations previously used to give a log odds score for matching each type of secondary structure to each position taking into account the level of reliability as given in the '`Rel sec`' line of the output file. Calculation of significance is performed using a Poisson fit method since this was known to work well for other sequence based searches.

A comparison of the old fixed reliabilities and new variable reliabilities shows a marked improvement in sensitivity with distant relationships despite the prediction being incorrect in places.

Figure 4.2 shows a comparison of the histograms for CASP2 Target `t0031` (see Chapter 5) using the same parameters but fixed and variable scoring. Although the top hit is the same (Figure 5.16) in both cases the fixed scoring scheme does not discriminate well between this and the noise. The variable scoring scheme has

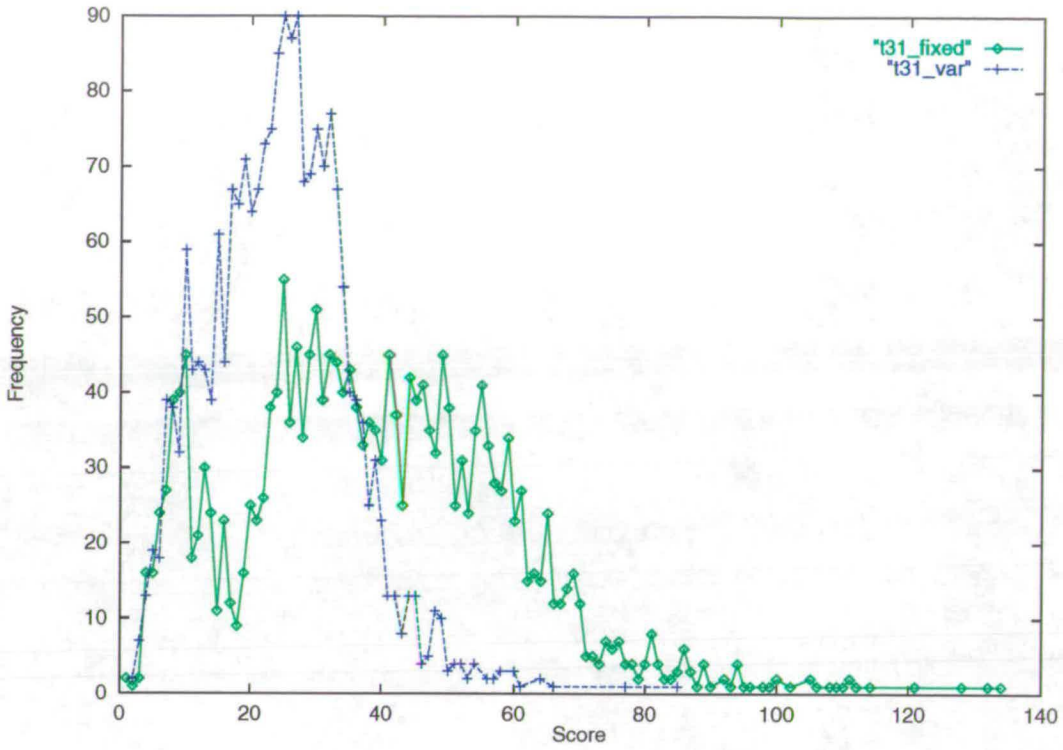


Figure 4.2: *Score distribution for fixed and variable scoring shemes using CASP2 target* `t0031`.

successfully reduced the noise by taking account of the less reliable parts of the prediction. This result is particularly impressive because the sequence identity for the top hit is less than 16% and the structure is predominantly β strands which tend to be poorly predicted.

4.3 Database development

In order to test the algorithm a database was needed. The simplest method of building one was to take the DSSP (Kabsch & Sander, 1983) database and make a non-redundant version. This reduction was achieved by first eliminating any sequences which were exact duplicates, this reduced the number of sequences by almost a half. Next it was necessary to do a comparison of all sequences to remove any which were clearly mutants of ones already seen. In this case, where sequences were greater than 90% identical, the shorter of the two was removed.

This resulted in another halving of the database with negligible loss of information. The initial size of the database was 7955 sequences and the filtered database was 2133 sequences.

In addition to this initial database, a larger one was derived from HSSP (Sander & Schneider, 1991) by cross referencing with the non-redundant DSSP set, and eliminating any high identity homologues from the new set. This database of homologues and real structures consisted of just 15249 sequences reduced from 44862, an easily searchable set for a Smith & Waterman algorithm. The advantage of the latter database is that there is a greater chance that a sequence in that set will be more closely homologous to a given query than if the sequences from only known structures were to be used. In addition, the use of a 'stepping stone' sequence should ensure a better quality of alignment more suited to homology modelling.

A good example of this is seen if CASP2 Target `t0031` is searched against this new database. While the search against the normal `sssdb` found two significant hits, this new search has the first eleven as significant. The alignment (Figure 4.3) has higher sequence identity than the alignment seen in Figure 5.16 suggesting that this sequence is more closely related to the target sequence than the sequence with known structure (Note that this and all subsequent alignments have been processed from the original `sss_align` output to show the secondary structures and sequence similarity more clearly).

4.4 Additional features

In addition to being useful for predicting structures, the program is designed to be used as a primary search tool. For this reason it can be used to search databases in `FASTA` and SwissProt/EMBL format in addition to `SSSDB` and `PHD` format. All formats can be searched against others and where either have no secondary structure information only PAM table scores are used. Since both query and database entries can be in any of the known formats it is possible to use the

Figure 4.3: *Alignment of* `t0031` *against the extended homologues database.*

program as a one-on-one alignment tool. This can be especially useful if two sequences are thought to be related distantly because PHD predictions can be made for both and an alignment made using this extra information.

For example, it was thought that the recently published crystal structure of firefly luciferase (Conti *et al.* 1996) was similar to a domain of cyclosporin synthetase (Malcom D. Walkinshaw, personal communication). Cyclosporin synthetase is the largest known single chain polypeptide consisting of 15281 amino acids (Weber *et al.* 1994). The protein is produced by the fungus *Tolypocladium inflatum Gams* for the non-ribosomal synthesis of the undecapeptide cyclosporin A (Dittmann *et al.* 1994). The sequence is made up of 11 homologous domains. These are responsible for the adenylation, thioesterification and (for seven of the amino acids) N-methylation, of each amino acid in the product. Homology between the aminoadenylation domain and the firefly luciferase was observed. Unfortunately the coordinates were not available. A solution was to predict the secondary structures for luciferase and and a domain of cyclosporin synthetase. Then `sss_align` was used to align the pair (Figure 4.4). The added secondary structure information gives a clearer idea of how the two domains are similar structurally. The strong sequence identity at residues 199–208 in luciferase (`Db`) and 636–645 of cyclosporin synthetase (`Qy`) is thought to be involved in ATP binding as this sequence of invariant residues is seen in a wide range of acyl adenylation proteins. It is found in all 11 domains.

The ability to output the sorted histogram of results for plotting as seen in Figure 4.2 was added to assist in analysing the effects of applying different parameter settings to the program. Suboptimal alignment output aids the identification of divergent duplicates of regions of sequences.

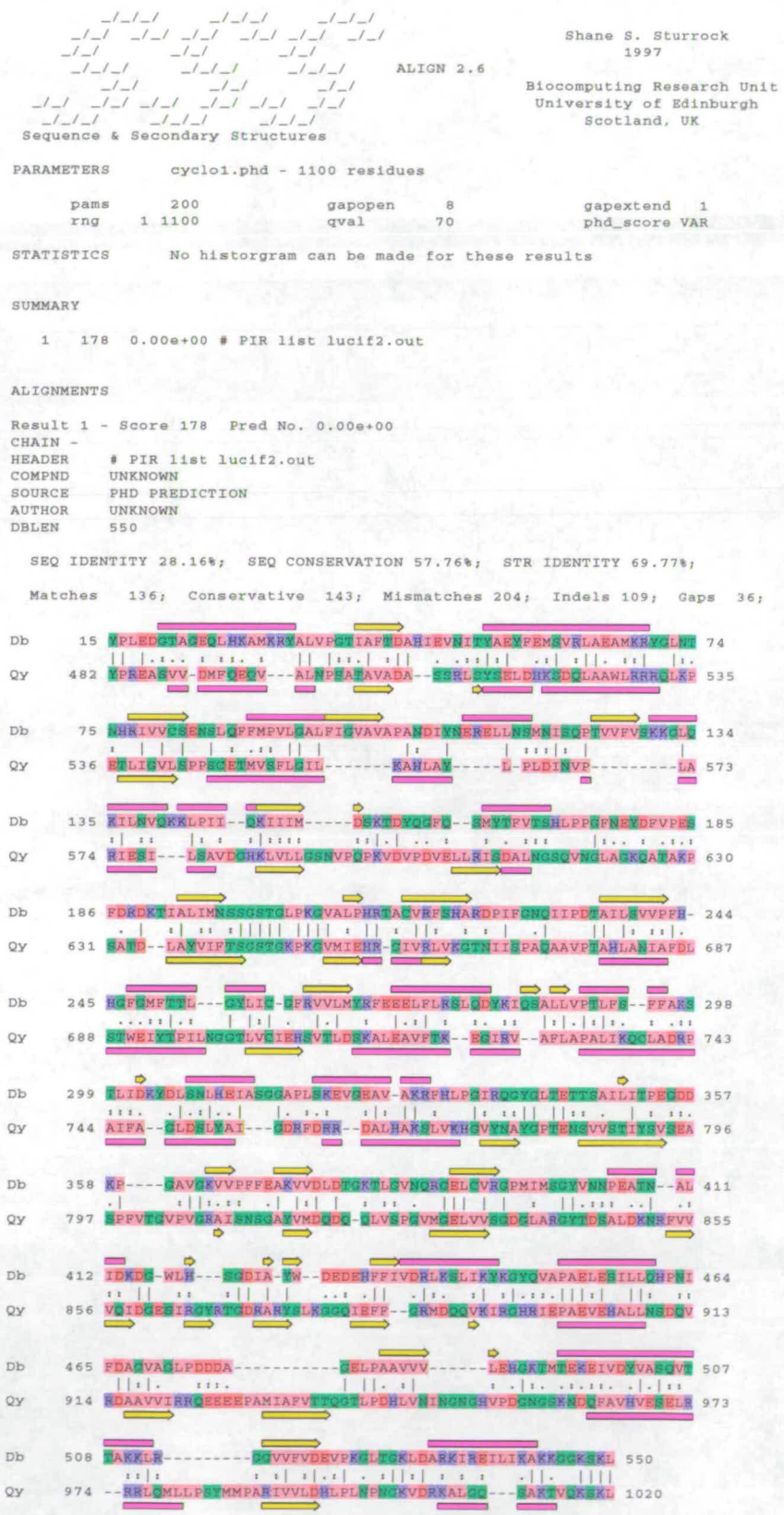


Figure 4.4: *Alignment of two sequences with predicted secondary structures.*

4.5 Summary

The program developed from an experimental testbed for the new scoring schemes to a full sequence analysis tool as a result of its use in the CASP2 homology challenge. It was necessary to make it fully functional so that Dr. A.F.W. Coulson could use it in the blind trial with ease. This trial was needed to verify that it would perform as expected. The next chapter describes the results.

Chapter 5

CASP2 results

5.1 Introduction

'Critical Assessment of techniques for protein Structure Prediction' (CASP2) is an international exercise to evaluate structure prediction methods. The sequences of about-to-be-solved protein structures were provided (from February 1996) by the CASP2 organisers as targets for blind prediction, in four categories:

- Comparative Modelling — A new sequence which has a high degree of sequence similarity to another sequence with a known structure, is modified to match the new sequence predicting the conformation of sidechains and loops so that the quality of the resulting model could be assessed.
- Fold Recognition — A new sequence which has no identifiable sequence similarity to any known structure is compared using more sensitive methods to recognise that the sequence belongs to a previously observed structural family.
- Ab initio Prediction — A new structure is to be predicted without the aid of a known structural analogue.
- Docking — The target is a complex for which the structure of isolated components are already known.

Although `sss_align` does not fall into the typical 'fold threading' group of programs it was decided to try and see if similar results could be achieved. For this reason we participated in the fold recognition part of the exercise where 22 targets were presented for prediction. The program runs and submission of predictions were performed by Dr. A.F.W. Coulson.

5.2 Submissions

Multiple submissions were made for each target because no development experiments had been done at the time in order to test the program under a wide range of conditions.

Of the 22 targets in the category predictions were submitted for 18. Those that were missed were the result of missing the deadlines for submission of a result.

For each target a grid search was performed. The QVAL range was 0, 5, 10, 30, 50, 70, 90, 95 and 100. The PAM range was 40, 80, 150, 250, 350. This grid search was performed with appropriate gap penalties for each PAM table, first with fixed gaps and then with affine gaps. For each target the grid search expected frequency and top hit PDB entry ID with gaps optimised for discrimination.

Each submission was made in 'CASP2' format which was designed to aid the evaluation of each submission.

An example 'CASP2' submission is shown in Figure 5.1 and can be compared with the equivalent normal `sss_align` output for the same sequence in Figure 5.16. The format includes information on how likely the program thinks the alignment is real (TSCORE), what segments of two sequences are aligned (TALIGN) and whether only part of the query was used (SEQSUB). By using this standard format for all entries it became possible to use an automatic evaluation process. Each entry was then compared against the set of similar structures according to various structure comparison programs. From this it was possible to rank hits by RMS deviation or residue shift (how far from the optimal structural alignment the submission is) for example.

```
PFRMAT FRV1
REMARK ============================================================================
REMARK                Shane S. Sturrock, BRU 1997
REMARK                 _/_/_/     _/_/_/      _/_/_/                 PARAMETERS
REMARK               _/_/  _/_/ _/_/  _/_/ _/_/  _/_/
REMARK              _/_/        _/_/        _/_/                     PAMS        250
REMARK               _/_/_/     _/_/_/      _/_/_/     ALIGN 2.6    GAPOPEN       8
REMARK                  _/_/        _/_/        _/_/                 GAPEXTEND     1
REMARK           _/_/  _/_/ _/_/  _/_/ _/_/  _/_/                    RNG     71  280
REMARK            _/_/_/      _/_/_/      _/_/_/                     QVAL         60
REMARK         Sequence & Secondary Structures                       PHD_SCORE   VAR
REMARK
REMARK Results for the target t0031, compared against a database of sequences
REMARK derived from PDB DSSP files and filtered to remove redundant identical
REMARK sequences.  Alignments are based on standard Smith Waterman best local
REMARK similarity with the addition of secondary structure prediction.
REMARK ============================================================================
AUTHOR 1253-6817-4864 A.F.W.Coulson, BRU,
AUTHOR 1253-6817-4864 University of Edinburgh, UK
REMARK ============================================================================
SEQRES t0031 MNNSKIISKVLLSLSLFTVGASAFVIQDELMQKNHAKAEVSAEEIKKHEEKWNKYYGVNA
SEQRES t0031 FNLPKELFSKVDEKDRQKYPYNTIGNVFVKGQTSATGVLIGKNTVLTNRHIAKFANGDPS
SEQRES t0031 KVSFRPSINTDDNGNTETPYGEYEVKEILQEPFGAGVDLALIRLKPDQNGVSLGDKISPA
SEQRES t0031 KIGTSNDLKDGDKLELIGYPFDHKVNQMHRSEIELTTLSRGLRYYGFTVPGNSGSGIFNS
SEQRES t0031 NGELVGIHSSKVSHLDREHQINYGVGIGNYVKRIINEKNE
REMARK ============================================================================
TSCORE t0031 1  5.854e+00   1PPF E 0
TSCORE t0031 1  1.504e+00   1TRY - 0
TSCORE t0031 1  2.485e-01   1BMA A 0
TSCORE t0031 1  1.809e-01   1SGT - 0
REMARK ============================================================================
TALIGN t0031 1    71    124        1PPF E 0       17     68       1.0       1
TALIGN t0031 1   129    156        1PPF E 0       69     97       1.0       1
TALIGN t0031 1   157    164        1PPF E 0      101    108       1.0       1
TALIGN t0031 1   168    184        1PPF E 0      109    125       1.0       1
TALIGN t0031 1   185    217        1PPF E 0      127    160       1.0       1
TALIGN t0031 1   218    246        1PPF E 0      166    201       1.0       1
TALIGN t0031 1   247    258        1PPF E 0      216    226       1.0       1
TALIGN t0031 1   260    276        1PPF E 0      227    243       1.0       1
REMARK ============================================================================
SEQSUB t0031 1    71    280
REMARK ============================================================================
```

Figure 5.1: *Example CASP2 format submission for* t0031.

Target	Protein	Residues	Fold	sss_align correct?	Model structure
t0004	Polyribonucleotide nuclotidyl transferase	84	Cold shock DNA-binding domain	Yes	1CSP
t0014	3-Dehydroquinase	252	TIM Barrel	Yes	1BGL
t0020	Ferrochelatase	320	Periplasmic binding-protein	No	1TLF
t0022	L-fucose isomerase	591	*N*-terminal 175 residues ferredoxin-like fold	No	2FX2
t0031	Exfoliative toxin A	242	Trypsin-like serine protease	Yes	1PPF
t0038	CBDN1	152	Galactose-binding domain	No submission	*

Table 5.1: *List of identifiable folds in CASP2 evaluation of Fold Recognition (* Since no submission was made the data on what was the correct fold was not provided to our group).*

5.3 Results

By the time of the evaluation 5 structures had not been solved and most of the remainder represented new folds which could not be predicted by analogy. The consensus of the structure alignment programs was that a known fold represented part or all of the structure in 6 cases. The results in these cases are shown for `sss_align` in Table 5.1. No prediction was submitted for `t0038`. The following sections will discuss each result.

```
                1                                                    50
predict_h148 AEIEVGRVYT GKVTRIVDFG AFVAIGGGKE GLVHISQIAD KRVEKVTDYL
pnp_ecoli    AEIEVGRVYT GKVTRIVDFG AFVAIGGGKE GLVHISQIAD KRVEKVTDYL
pnp_pholu    AEIEVGRIYA GKVTRIVDFG AFVAIGGGKE GLVHISQIAD KRVEKVADYL
pnp_haein    AEVEAGVIYK GKVTRLADFG AFVAIVGNKE GLVHISQIAE ERVEKVSDYL
rs1h_bacsu   QSLEVGSVLD GKVQRLTDFG AFVDIGG.ID GLVHISQLSH SHVEKPSDVV
yabr_bacsu   MSIEVGSKLQ GKITGITNFG AFVELPGGST GLVHISEVAD NYVKDINDHL
rs1_rhime    AKYPVGKKIS GTVTNITDYG AFVELEPGIE GLIHISEMst KKNVHPGKIL
rs1_mycle    .THAIGQIVP GKVTKLVPFG AFVRVEEGIE GLVHISELAE RHVEVPDQVV
rr1_spiol    AQLGIGSVVT GTVQSLKPYG AFIDIGG.IN GLLHVSQISH DRVSDIATVL
yhgf_ecoli   NDLQPGMILE GAVTNVTNFG AFVDIGVHQD GLVHISSLSN KFVEDPHTVV
pr22_yeast   ....LHKVYE GKVRNITTFG CFVQIFGTrd GLVHISEMSD QRTLDPHDVV
rs1_synp6    NRLEVGEVVV GAVRGIKPYG AFIDIGG.VS GLLHISEISH DHIETPHSVF
rs1_prosp    ENLQEGMEVK GIVKNLTDYG AFVDLGG.VD GLLHITDMAW KRVKHPSEIV
rs1_ecoli    ENLQEGMEVK GIVKNLTDYG AFVDLGG.VD GLLHITDMAW KRVKHPSEIV
rpoe_sulac   ....IHEVIE GEVSQVDNYG VYVNMGP.VD GLVHISQITD DNleKSKKSI
rs1_chltr    SEVQPGAILK GTVVDISKDF VVVDVGLKSE GVIPMSEFID S.....SEGL
rne_haein    HEQKKANIYK GKITRVEPsa AFVDYGAERH GFLPLKEIAR EYFpnIRDIL
rne_ecoli    HEQKKANIYK GKITRIEPsa AFVDYGAERH GFLPLKEIAR EYFpnIKDVL

                51                                    84
predict_h148 QMGQEVPVKV LEVDRQGRIR LSIKEATEQS QPAA
pnp_ecoli    QMGQEVPVKV LEVDRQGRIR LSIKEATEQS QPAA
pnp_pholu    QVGQETSVKV LEIDRQGRVR LSIKEATAGT AVEE
pnp_haein    QVGQEVNVKV VEIDRQGRIR LTMKDLAPKQ ETE.
rs1h_bacsu   EEGQEVKVKV LSVDRdeRIS LSIKDTLP.. ....
yabr_bacsu   KVGDQVEVKV INVEKDGKIG LSIKKAKDRP QARP
rs1_rhime    STSQEVDVVV LEVDPtrRIS LGLKQTLENP WQA.
rs1_mycle    AVGDDAMVKV IDIDLerRIS LSLKA..... ....
rr1_spiol    QPGDTLKVMI LSHDRegRVS LSTKKLEP.. ....
yhgf_ecoli   KAGDIVKVKV LEVDLqkRIA LTMRLDEQPG ETNA
pr22_yeast   RQGQHIFVEV IKIQNNGKIS LSMKNIDQHS ....
rs1_synp6    NVNDEVKVMI IDLDAegRIS LSTKQLEPE. ....
rs1_prosp    NVGDEITVKV LKFDRetRVS LGLKQLGEDP WVA.
rs1_ecoli    NVGDEITVKV LKFDRetRVS LGLKQLGEDP WVA.
rpoe_sulac   TKGDRVRAMI IssGRLPRIA LTMKQP.... ....
rs1_chltr    SVGAEVEVYL DqeDEEGKVV LSREKATRQR Q...
rne_haein    VEGQEVIVQV NKEERGNK.. .......... ....
rne_ecoli    REGQEVIVQI DKEERGNK.. .......... ....
```

Figure 5.2: *Multiple alignment of* t0004 *and its close homologues as used as input for* PHD *secondary structure prediction.*

5.3.1 Target t0004

Target t0004 was the least taxing of the sequences presented. Searches with sequence similarity alone gave the correct sequence as the top hit with some scoring tables. However, the hit was not classed as significant by the expected frequency. The secondary structure prediction was quite accurate because there were plenty of homologues in the SwissProt database (Figures 5.2, 5.3). Table 5.2 shows the grid search results for this target.

As can be seen from Table 5.2 and Figure 5.3 the inclusion of very little secondary structure information was enough to make the result significant over the noise.

Figure 5.3 shows the alignment generated by `sss_align` the best result of the

Q/PAMS	40	80	110	150	250	350
0	4GPB	4GPB	4GPB	4GPB	4GPB	4GPB
	2.36e+00	2.37e+00	2.65e+00	2.32e+00	3.29e+00	2.35e+00
5	4GPB	4GPB	4GPB	4GPB	4GPB	4GPB
	1.70e+00	1.79e+00	2.02e+00	1.94e+00	2.70e+00	2.05e+00
10	4GPB	4GPB	4GPB	4GPB	4GPB	4GPB
	1.34e+00	1.52e+00	1.70e+00	1.78e+00	2.41e+00	2.02e+00
30	1CSS	1BUC	1BUC	1BUC	1BUC	4GPB
	6.87e+00	6.92e+00	2.82e+00	4.60e+00	3.13e+00	4.95e+00
50	1RNR	1CDK	1CDK	1BUC	1BUC	1BUC
	3.10e-01	1.55e+00	2.44e+00	1.53e+00	1.22e+00	3.28e+00
70	2YHX	1ATP	1CSP	1CSP	1CSP	1CSP
	1.10e+01	5.02e+00	4.67e+00	5.35e-01	1.15e+00	1.41e+01
90	1TYT	1MJC	1CSP	1CSP	1CSP	1HNV
	1.66e+01	2.96e+01	2.34e+00	1.95e-01	5.68e+00	2.39e+01
95	1ACF	1GEU	1CSP	1CSP	1OAC	1HNV
	1.42e+01	1.55e+01	3.88e+00	7.45e-01	1.68e+01	4.38e+01
100	1ACF	1MLA	1CSP	4CSP	1OAC	1RTJ
	5.22e+00	5.66e+00	1.15e+01	7.57e+00	4.55e+01	1.01e+02

Table 5.2: *PDB identifier and expected frequency for each search with* `t0004`.

Figure 5.3: *Alignment of* 1CSP (Db) *and* t0004 (Qy).

survey (Table 5.2). There is an high degree of sequence similarity and agreement in secondary structure between the target and 1CSP which is a cold shock protein from *E.coli*. Figure 5.4 compares the two structures which are both OB-fold barrels.

5.3.2 Target t0014

Target t0014 had a more limited number of homologues (Figure 5.5) than t0004 but with a good range of variation it was expected that the prediction would be fairly accurate. The results for the grid search are shown in Table 5.3.

1BGL has a region of structure which is aligned by sss_align under a wide range of conditions.

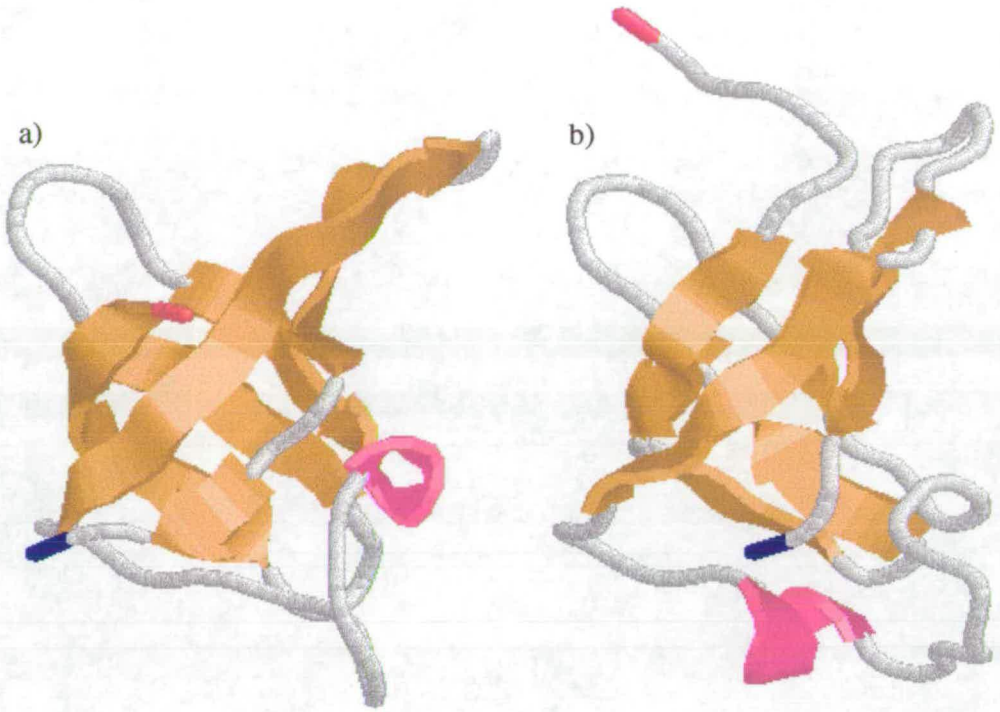


Figure 5.4: *Comparison of the structures of* a) `1CSP` *and* b) `t0004`

Apparently significant results in the lower left of the table were short, highly-matched regions which could not be extended over a useful sequence length.

Examination of the best alignment (Figure 5.6) shows a very low sequence identity but high agreement in structure to `1BGL`, β-galactosidase. The alignment covers more than half of the sequence of `t0014` and the secondary structure prediction makes it plausible that the second half of the structure duplicates the fold of the first half. This would produce a TIM barrel for the overall structure and publication of the structure revealed this to be correct.

This shows the ability of the local similarity algorithm used in `sss_align` to pick out homologous regions from within larger structures (Figures 5.7, 5.8 & 5.9).

5.3.3 Target t0020

Target `t0020` had a good range of homologues (Figure 5.10) in the database and was expected to have an accurate prediction. The grid search results are shown in Table 5.4.

```
                1                                                   50
predict_h172    MKTVTVKNLI IGEGMPKIIV SLMGRDINSV KAEALAYREA TFDILEWRVD
arod_salti      MKTVTVKNLI IGEGMPKIIV SLMGRDINSV KAEALAYREA TFDILEWRVD
arod_ecoli      MKTVTVKDLV IGTGAPKIIV SLMAKDIASV KSEALAYREA DFDILEWRVD
arod_bacsu      MNVLTIKGVS IGEGMPKIII PLMGKTEKQI LNEAEAVKLL NPDIVEWRVD
ebsd_entfa      MKPVIVKNVR IGEGNPKIVV PIVAPTAEDI LAEATASQTL DCDLVEWRLD
acu20284_1,_cut .STYVVKNLN IGDLPVKTLV PITAKTREQA LAQAKVIAek DADIAEFRID
S46210,_cut     .......... ..SGVRKMEG EAMTRNETLI CAPIMADTvs GADLVEVRLD
S486918,_cut    .......... .GETVEKMVV DIQKAKLNGA .......... ..DLVEIRLD

                51                                                  100
predict_h172    HFMDIASTQS VLTAARVIRD AMPDIPLLFT FRSAKEGGEQ TITTQHYLTL
arod_salti      HFMDIASTQS VLTAARVIRD AMPDIPLLFT FRSAKEGGEQ TITTQHYLTL
arod_ecoli      HYADLSNVES VMAAAKILRE TMPEKPLLFT FRSAKEGGEQ AISTEAYIAL
arod_bacsu      VFEKANDREA VTKLISKLRK SLEDKLFLFT FRTHKEGGSM EMDESSYLAL
ebsd_entfa      YYENVADFSD VCNLSQQVME RLGQKPLLLT FRTQKEGGEM AFSEENYFAL
acu20284_1,_cut LLEFASDTKK VIALGQELNQ ILKDKPLLAT IRTSNEGGKL KVTDQEYEKI
S46210,_cut     SLKSFNPQSD IDTII..... KQSPLPTLFT YRPTWEGGQY AGDEVSRLDA
S486918,_cut    SLSTFNPHQD LNTFI....Q QHHSLPFLFT YRPIWEGGKY DGDENRRLDA

                101                                                 150
predict_h172    NRAAIDSGLV DMIDLELFTG DADVKATVDY AHAHNVYVVM SNHDFHQTPS
arod_salti      NRAAIDSGLV DMIDLELFTG DADVKATVDY AHAHNVYVVM SNHDFHQTPS
arod_ecoli      NRAAIDSGLV DMIDLELFTG DDQVKETVAY AHAHDVKVVM SNHDFHKTPE
arod_bacsu      LESAIQTKDI DLIDIELFSG DANVKALVSL AEENNVYVVM SNHDFEKTPV
ebsd_entfa      YHELVKKGAL DLLDIELFAN PLAADTLIHE AKKAGIKIVL CNHDFQKTPS
acu20284_1,_cut YSEYLKKPFM QLLDIEMFRD QAAVAKLTKL AHQKKVLVVM SNHDFDKTPS
S46210,_cut     LRVAMELG.A DYIDVELKAI D.EFNTALHG NKSAKCKVIV SSHNYDNTPS
S486918,_cut    LRLAVELG.A DYVDIELKVA HEFYDSIRGK MF.NKTKVIV SSHNYQYTPS

                151                                                 200
predict_h172    AEEMVSRLRK MQALGADIPK IAVMPQSKHD VLTLLTATLE MQQHYADRPV
arod_salti      AEEMVSRLRK MQALGADIPK IAVMPQSKHD VLTLLTATLE MQQHYADRPV
arod_ecoli      AEEIIARLRK MQSFDADIPK IALMPQSTSD VLTLLAATLE MQEQYADRPI
arod_bacsu      KDEIISRLRK MQDLGAHIPK MAVMPNDTGD LLTLLDATYT MKTIYADRPI
ebsd_entfa      QEEIVARLRQ MQMRQADICK IAVMPQDATD VLTLLSATNE MYTHYASVPI
acu20284_1,_cut EQEIVSRLLK QDQMGADILK IAVMPKSKQD VFTLMNATLK VSEQST.KPL
S46210,_cut     SEELGNLVAR IQASGADIVK FATTALDIMD VARVFQITVH SQ.....VPI
S486918,_cut    VEDLGDLVAR IQATGADIVK IATTAVEITD VARMFQILVH SQ.....VPF

                201                                                 250
predict_h172    ITMSMAKEGV ISRLAGEVFG SAATFGAVKQ ASAPGQIAVN DLRSVLMILH
arod_salti      ITMSMAKEGV ISRLAGEVFG SAATFGAVKQ ASAPGQIAVN DLRSVLMILH
arod_ecoli      ITMSMAKTGV ISRLAGEVFG SAATFGAVKK ASAPGQISVN DLRTVLTILH
arod_bacsu      ITMSMAATGL ISRLSGEVFG SACTFGAGEE ASAPGQIPVS ELRSVLDILH
ebsd_entfa      VTMSMGQLGM ISRVTGQLFG SALTFGSAQQ ASAPGQLSVQ VLRNYLKTF.
acu20284_1,_cut LTMSMGRLGT ISRIATANMG GSLSFGMIGE ASAPGQIDVT ALKQFLKTVQ
S46210,_cut     IAMVMGEKGL MSRILCPKFG GYLTFGTLek VSAPGQPTIK DLLNI.....
S486918,_cut    IGLVMGDRGL VSRVLCAKFG GYLTFGTLev VSAPGQPTIK DLLHL.....

                252
predict_h172    NA
arod_salti      NA
arod_ecoli      QA
arod_bacsu      K.
ebsd_entfa      ..
acu20284_1,_cut ..
S46210,_cut     ..
S486918,_cut    ..
```

Figure 5.5: *Multiple alignment of* `t0014` *and its close homologues as used as input for* PHD *secondary structure prediction.*

Q/PAMS	40	80	110	150	250	350
0	1BGL	1BGL	1BGL	1BGL	1BGL	1BGL
	4.66e+01	4.33e+01	5.97e+01	4.64e+01	6.89e+01	3.80e+01
5	1BGL	1BGL	1BGL	1BGL	1BGL	1BGL
	2.79e+01	2.78e+01	4.62e+01	3.34e+01	6.81e+01	2.97e+01
10	1BGL	1BGL	1BGL	1BGL	1BGL	1BGL
	1.48e+01	1.68e+01	3.12e+01	2.36e+01	5.71e+01	2.35e+01
30	1BGL	1BGL	1BGL	1BGL	1BGL	1BGL
	6.70e-01	1.10e+00	3.88e+00	4.15e+00	1.95e+01	1.14e+01
50	1BGL	1SUC	1BGL	1BGL	1BGL	1LTP
	8.13e+00	7.30e+00	7.20e+00	1.07e+01	1.61e+01	1.47e+01
70	1PRT	1SEL	1SCN	1SEL	1OPR	1LTP
	2.59e+00	2.86e+00	1.14e+00	2.42e+00	1.10e+01	4.08e+01
90	1PTO	1SEL	1SEL	1DRK	1ADD	1IDC
	3.40e-01	5.05e+00	4.51e+00	6.29e+00	4.58e+01	1.25e+02
95	1PTO	1PHP	1LIN	1CDE	1IDF	P1MIO
	5.26e-01	6.42e+00	2.16e+00	1.54e+01	7.57e+01	1.61e+02
100	1PGQ	1PHP	1CDE	1BVP	1IDF	1PTH
	3.73e-01	3.64e+00	1.09e+00	1.02e+01	1.05e+02	1.99e+02

Table 5.3: *PDB identifier and expected frequency for each search with* `t0014`.

Figure 5.6: *Alignment of* 1BGL (Db) *and* t0014 (Qy).

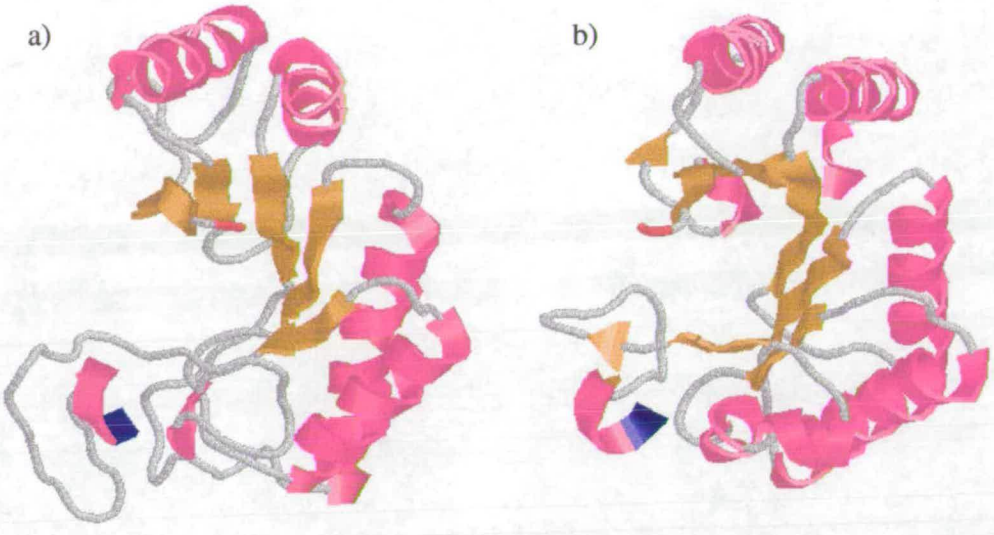


Figure 5.7: *Comparison of the aligned regions of structures of* a) `1BGL` *and* b) `t0014`

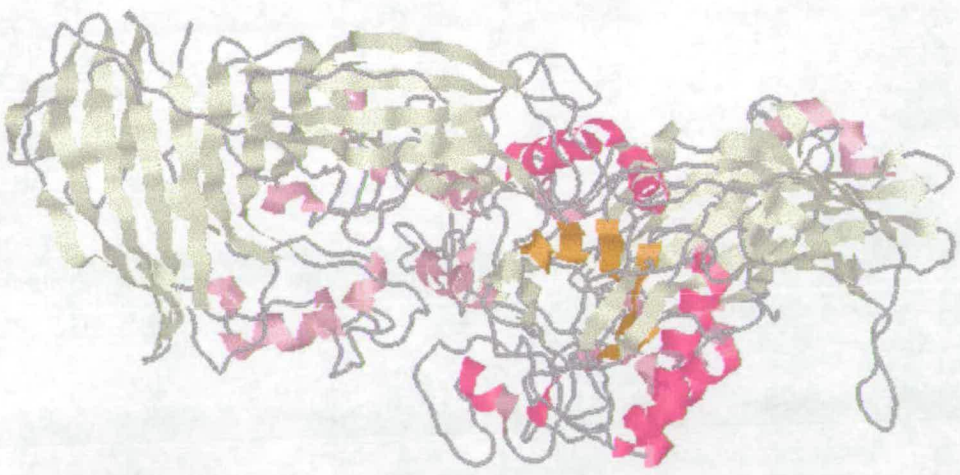


Figure 5.8: *Full structure of* `1BGL` *chain A highlighting the region of homology to* `t0014`.

Q/PAMS	40	80	110	150	250	350
0	1ASZ	1ASZ	1ASZ	1ASZ	1ASZ	1ASZ
	8.40e+01	8.28e+01	8.91e+01	8.37e+01	9.80e+01	8.27e+01
5	1BGL	1BGL	1ASZ	1BGL	1ASZ	1BGL
	6.50e+01	6.74e+01	8.26e+01	7.60e+01	9.28e+01	7.44e+01
10	1BGL	1BGL	1BGL	1BGL	1ASZ	1BGL
	3.81e+01	4.43e+01	6.42e+01	5.73e+01	8.80e+01	6.49e+01
30	1BGL	1RAI	1CHR	1CHR	1CHR	1CHR
	9.02e+00	1.02e+01	1.60e+01	1.69e+01	4.13e+01	3.16e+01
50	2MTA	1LXA	8API	8API	1GPA	1DPG
	1.47e+00	7.44e+00	4.74e+00	4.33e+00	3.90e+00	7.79e+00
70	2APD	2APD	1FPS	1FPS	7GPB	1DPG
	2.04e-01	1.51e+00	2.56e+00	5.31e+00	9.88e-01	1.14e+01
90	1FPS	1FPS	1FPS	1FPS	1GLN	1GLN
	1.33e+00	5.17e-01	9.18e-01	2.28e+00	2.37e+00	8.28e+01
95	1FPS	1FPS	1FPS	1EDT	1GLN	1GPA
	2.47e+00	7.63e-01	7.63e-01	2.09e+00	4.33e+01	9.72e+01
100	1STD	1PNG	1YMA	1IDM	1GLN	1PRJ
	2.99e+00	1.46e+00	2.28e+00	3.57e+00	8.48e+01	1.26e+02

Table 5.4: *PDB identifier and expected frequency for each search with* `t0020`.

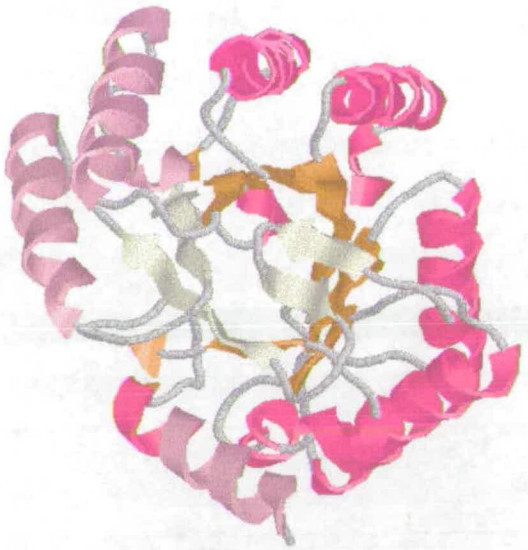


Figure 5.9: *Full structure of* t0014 *highlighting the region of homology to* 1BGL *chain A.*

According to the structure comparison programs the fold is most similar to the 'periplasmic binding protein-like I' fold. The released structure of t0020 is made up of two four-stranded parallel β sheets, with sheet order 2134. The axes of the sheets are roughly parallel, and the C-terminal sides of the sheets are not quite pointed together. The closest structure to this found by the structure comparison programs was 1TLF, a transcription regulation protein from *E.coli*. This is shown with t0020 in Figure 5.12. In 1TLF the two sheets have six strands and the axes are almost at right angles. Strand order is 213456 which is the same as t0020 except for the extra two strands.

The secondary structure prediction for this target is very close to that found in the released structure (Figure 5.11). Despite this and even knowing the correct structure it is not possible to get the correct alignment with sss_align. The alignment shows 13.95% sequence identity and 85.12% structure identity but in a search this hit does not look significant appearing far into the output list. The problem appears to be a combination of the regular secondary structure and low sequence similarity. The structural alignment only has about 6% sequence identity which suggests that these two structures either have no common ancestor or it is so distant that it is unrecognisable by sequence alignment.

```
                    1                                                  50
predict_e140        RKKMGLLVMA YGTPYKEEDI ERYYTHIRRG RKPEPEMLQD LKDRYEAIGG
C47045 C47045 f     RKKMGLLVMA YGTPYKEEDI ERYYTHIRRG RKPEPEMLQD LKDRYEAIGG
A54125 A54125 f     .......... .......... .......... .......... .KEGYAAIGG
A37972 A37972;      .......... .......... .......... .......... .......IGG
A36403 A36403;      .......... .......... .......... .......... IQEQYRRIGG
S16118 S16118;      .......... .......... .......... .......... .........G
JC2266 JC2266;      .......... .......... .......... .......... .........G
IBBYFC A35190;      .......... .......... .......... .......... ...QYREIGG

                    51                                                 100
predict_e140        ISPLAQITEQ QAHNLEQHLN EIQDEITFKA YIGLKHIEPF IEDAVAEMHK
C47045 C47045 f     ISPLAQITEQ QAHNLEQHLN EIQDEITFKA YIGLKHIEPF IEDAVAEMHK
A54125 A54125 f     GSPLRKITDE QADAIKMSLQ A..KNIAANV YVGMRYWYPF TEEAVQQIKK
A37972 A37972;      GSPIKMWTSK QGEGMVKLLD ELsaTAPHKY YIGFRYVHPL TEEAIEEMER
A36403 A36403;      GSPIKIWTSK QGEGMVKLLD ELsnTAPHKY YIGFRYVHPL TEEAIEEMER
S16118 S16118;      GSPLMVYSRQ QQQALAQRLP E......MPV ALGMSYGSPS LESAVDELLA
JC2266 JC2266;      GSPLMVYSRQ QQQALAQRLP E......MPV ALGMSYGSPS LESAVDELLA
IBBYFC A35190;      GSPIRKWSEY QATEVCKILD KtpETAPHKP YVAFRYAKPL TAETYKQMLK

                    101                                                150
predict_e140        DGITEAVSIV LAPHFSTFSV QSYNKRAKEE AEKLGGLTIT SVESWYDEPK
C47045 C47045 f     DGITEAVSIV LAPHFSTFSV QSYNKRAKEE AEKLGGLTIT SVESWYDEPK
A54125 A54125 f     DKITRLVVLP LYPQYSISTT GSSIRVLQDL FRklAGVPVA IIKSWYQRRG
A37972 A37972;      DGLERAIAFT QYPQYSCSTT GSSLnrYYNE VGQKPTMKWS TIDRWPTHPL
A36403 A36403;      DGLERAIAFT QYPQYSCSTT GSSLnrYYNQ VGRKPTMKWS TIDRWPTHHL
S16118 S16118;      EHVDHIVVLP LYPQFSCSTv aVWDELARIL ARKRSIPGIS FIRDYADNHD
JC2266 JC2266;      EHVDHIVVLP LYPQFSCSTv aVWDELARIL ARKRSIPGIS FIRDYADNHD
IBBYFC A35190;      DGVKKAVAFS QYPHFSYSTT GSSINELWRQ IKAlrSISWS VIDRWPTNEG

                    151                                                200
predict_e140        FVTYWVDRVK ETYASMPEDE RENAMLIVSA HSLPEKIKEF GDPYPDQLHE
C47045 C47045 f     FVTYWVDRVK ETYASMPEDE RENAMLIVSA HSLPEKIKEF GDPYPDQLHE
A54125 A54125 f     YVNSMADLIE KELQTFS..D PKEVMIFFSA HGVpsYVENA GDPYQKQMEE
A37972 A37972;      LIQCFADHIL KELNHFPEEK RSEVVILFSA HSLPMSVVNR GDPYPQEVGA
A36403 A36403;      LIQCFADHIL KELDHFPLEK RSEVVILFSA HSLPMSVVNR GDPYPQEVSA
S16118 S16118;      YINALANSVR ASFAKHGEPD ....LLLLSY HGIPQRYADE GDDYPQRCRT
JC2266 JC2266;      YINALANSVR ASFAKHGEPD ....LLLLSY HGIPQRYADE GDDYPQRCRT
IBBYFC A35190;      LIKAFSENIT KKLQEFPQPV RDKVVLLFSA HSLPMDVVNT GDAYPAEVAA

                    201                                                250
predict_e140        SAKLIAEGAG VSEYAVGWQS EGNTPDPWLG PDVQDLTRDL FEQKGYQAFV
C47045 C47045 f     SAKLIAEGAG VSEYAVGWQS EGNTPDPWLG PDVQDLTRDL FEQKGYQAFV
A54125 A54125 f     CIDLIMerGV LNDHKLAYQS RV.GPVQWLK PYTDEVLVDL GK.SGVKSLL
A37972 A37972;      TvkVMEKLGY PNPYRLVWQS KV.GPVPWLG PQTDEAIKGL CE.RGRKNIL
A36403 A36403;      TvkVMERLEY CNPYRLVWQS KV.GPMPWLG PQTDESIKGL CE.RGRKNIL
S16118 S16118;      TTRELASALG MAPEKvtFQS RF.GREPWLM P.YTDETLKM LGEKGVGHIQ
JC2266 JC2266;      TTRELASALG MAPEKvtFQS RF.GREPWLM P.YTDETLKM LGEKGVGHIQ
IBBYFC A35190;      TvnIMQKLKF KNPYRLVWQS QV.GPKPWLG AQTAEIAEFL GP..KVDGLM

                    251                                                300
predict_e140        YVPVGFVADH LEVLYDNDYE CKVVTDDIGA SYYRPEMPNA KPEFIDALAT
C47045 C47045 f     YVPVGFVADH LEVLYDNDYE CKVVTDDIGA SYYRPEMPNA KPEFIDALAT
A54125 A54125 f     AVPVSFVSEH IETLEEIDME YRELALESge NWGRVPALGL TPSFITDLAD
A37972 A37972;      LVPIAFTSDH IETLYELDIE YSQVLAQkaE NIRRAESLNG NPLFSKALAD
A36403 A36403;      LVPIAFTSDH IETLYELDIE YSQVLAKEce NIRRAESLNG NPLFSKALAD
S16118 S16118;      VMCPGFAADC LETLEEIAEQ NREVFLGAGg kYEYIPALNA TPEHIEMMAN
JC2266 JC2266;      VMCPGFAADC LETLEEIAEQ NREVFLGAGg kYEYIPALNA TPEHIEMMAN
IBBYFC A35190;      FIPIAFTSDH IETLHEIDLG V.IGESEYKD KFKRCESLNG NQTFIEGMAD

                    301  308
predict_e140        VVLKKLGR
C47045 C47045 f     VVLKKLGR
A54125 A54125 f     AVIESL..
A37972 A37972;      LV......
A36403 A36403;      LV......
S16118 S16118;      LV......
JC2266 JC2266;      LV......
IBBYFC A35190;      LV......
```

Figure 5.10: *Multiple alignment of* t0020 *and its close homologues as used as input for* PHD *secondary structure prediction.*

Figure 5.11: *Alignment of real* (Db) *and predicted* (Qy) *structures of* t0020.

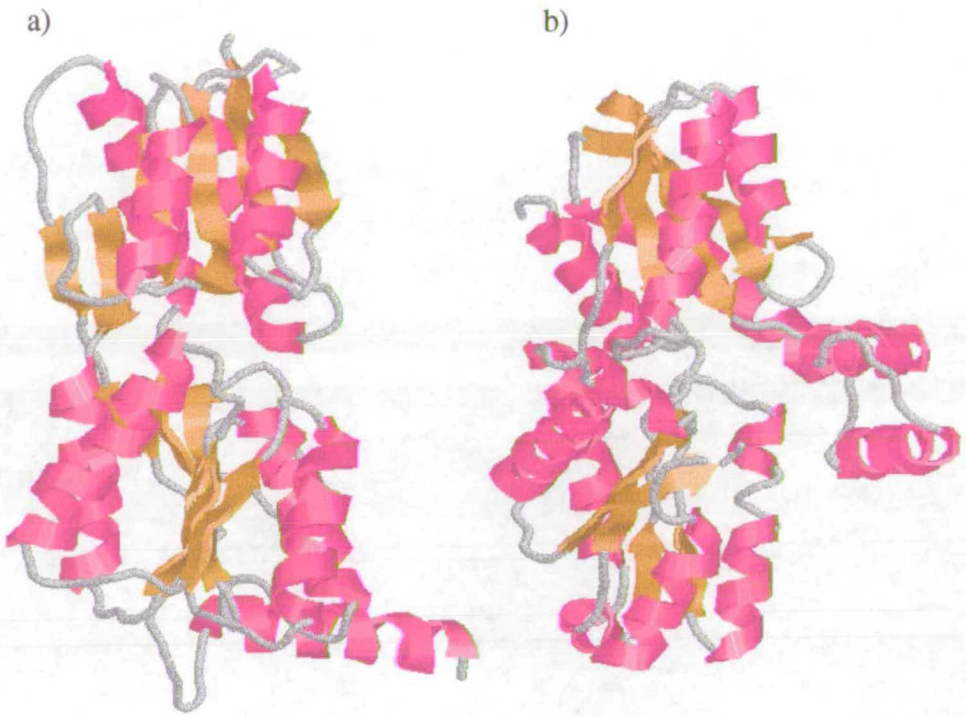


Figure 5.12: *Comparison of the structures of* a) `1TLF` *and* b) `t0020`

5.3.4 Target t0022

Target `t0022` had only one homologue (Figure 5.13) in the database and as such it was expected that the prediction would be very poor. There is no practical way to check the quality of the prediction because the released structure only contains $C\alpha$ atoms. The grid search results are in Table 5.5.

A couple of hits in the grid search look significant but neither is correct. The concensus of the structure alignment programs is that residues 1–175 of the structure are similar to a Ferredoxin like fold `2FX2`. As with `t0020` there is no recognisable sequence similarity. A comparison of the structures is shown in Figure 5.14.

5.3.5 Target t0031

Target `t0031` had a good multiple alignment (Figure 5.15) and was expected to have a good secondary structure prediction.

```
                    1                                                  50
predict_e275        MKKISLPKIG IRPVIDGRRM GVRESLEEQT MNMAKATAAL LTEKLRHACG
ECAE000364_2 AE     MKKISLPKIG IRPVIDGRRM GVRESLEEQT MNMAKATAAL LTEKLRHACG
HIU32743_10 U32     .......KIG IRPTIDGRRM GVRESLETQT IRMAQSVAQL LQTHIRHTDG

                    51                                                 100
predict_e275        AAVECVISDT CIAGMAEAAA CEEKFSSQNV GLTITVTPCW CYGSETIDMD
ECAE000364_2 AE     AAVECVISDT CIAGMAEAAA CEEKFSSQNV GLTITVTPCW CYGSETIDMD
HIU32743_10 U32     TFVECVVADS TIGGVAEAAA CADKFKRENV GLTITVTPCW CYGSETIDMD

                    101                                                150
predict_e275        PTRPKAIWGF NGTERPGAVY LAAALAAHSQ KGIPAFSIYG HDVQDADDTS
ECAE000364_2 AE     PTRPKAIWGF NGTERPGAVY LAAALAAHSQ KGIPAFSIYG HDVQDADDTS
HIU32743_10 U32     PHMPKAIWGF NGTERPGAVY LAAALAGHSQ LGLPAFSIYG TEVQEADDTN

                    151                                                200
predict_e275        IPADVEEKLL RFARAGLAVA SMKGKSYLSL GGVSMGIAGS IVDHNFFESW
ECAE000364_2 AE     IPADVEEKLL RFARAGLAVA SMKGKSYLSL GGVSMGIAGS IVDHNFFESW
HIU32743_10 U32     IPEDVKEKLL RFARAGLAVA SIRGKSYLSI GSVSMGIAGS IVNQAFFQEY

                    201                                                250
predict_e275        LGMKVQAVDM TELRRRIDQK IYDEAELEMA LAWADKNFRY GEDENNKQYQ
ECAE000364_2 AE     LGMKVQAVDM TELRRRIDQK IYDEAELEMA LAWADKNFRY GEDENNKQYQ
HIU32743_10 U32     LGMRNEYVDM MEIKRRLDRK IYDQEEVDLA LSWVKQYCKE GVDVNSLENQ

                    251                                                300
predict_e275        RNAEQSRAVL RESLLMAMCI RDMMQGNSKL ADIGRVEESL GYNAIAAGFQ
ECAE000364_2 AE     RNAEQSRAVL RESLLMAMCI RDMMQGNSKL ADIGRVEESL GYNAIAAGFQ
HIU32743_10 U32     RNAEERAELW ENVVKMTIIT RDLMVGNPKL ATLNYAEEAL GHNAIAAGFQ

                    301                                                350
predict_e275        GQRHWTDQYP NGDTAEAILN SSFDWNGVRE PFVVATENDS LNGVAMLMGH
ECAE000364_2 AE     GQRHWTDQYP NGDTAEAILN SSFDWNGVRE PFVVATENDS LNGVAMLMGH
HIU32743_10 U32     GQRHWTDHLP NGDFMEAMLN STYDWNGVRP PYILATENDS LNAIGMLFGH

                    351                                                400
predict_e275        QLTGTAQVFA DVRTYWSPEA IERVTGHKLD GLAEHGIIHL INSGSAALDG
ECAE000364_2 AE     QLTGTAQVFA DVRTYWSPEA IERVTGHKLD GLAEHGIIHL INSGSAALDG
HIU32743_10 U32     QLTGKAQIFA DVRTYWSQDS VERVTGWR.. ..PESGFIHL INSGSAALDG

                    401                                                450
predict_e275        SCKQRDSEGN PTMKPHWEIS QQEADACLAA TEWCPAIHEY FRGGGYSSRF
ECAE000364_2 AE     SCKQRDSEGN PTMKPHWEIS QQEADACLAA TEWCPAIHEY FRGGGYSSRF
HIU32743_10 U32     TGEHQDAQGN PTLKPAWDVT EEEAKRCLEN TRWCPAVHEY FRGGGLSSQF

                    451                                                500
predict_e275        LTEGGVPFTM TRVNIIKGLG PVLQIAEGWS VELPKDVHDI LNKRTNSTWP
ECAE000364_2 AE     LTEGGVPFTM TRVNIIKGLG PVLQIAEGWS VELPKDVHDI LNKRTNSTWP
HIU32743_10 U32     LTKGGIPFTI HRINLIKGLG PVLQIAEGWS IDLPQDVHNK LNQRTNETWP

                    501                                                550
predict_e275        TTWFAPRLTG KGPFTDVYSV MANWGANHGV LTIGHVGADF ITLASMLRIP
ECAE000364_2 AE     TTWFAPRLTG KGPFTDVYSV MANWGANHGV LTIGHVGADF ITLASMLRIP
HIU32743_10 U32     TTWFVPRLTG KGAFTDVYSV MANWGANHCV ATHGHVGADL ITLASMLRIP

                    551                                          591
predict_e275        VCMHNVEETK VYRPSAWAAH GMDIEGQDYR ACQNYGPLYK R
ECAE000364_2 AE     VCMHNVEETK VYRPSAWAAH GMDIEGQDYR ACQNYGPLYK R
HIU32743_10 U32     VCMHNVSEKN IFRPSAWNGF GQDKEGQDYR ACQNFGPLYK .
```

Figure 5.13: *Multiple alignment of* t0022 *and its close homologues as used as input for* PHD *secondary structure prediction.*

Q/PAMS	40	80	110	150	250	350
0	2TMD	2TMD	2TMD	2TMD	2TMD	2TMD
	1.80e+01	1.83e+01	1.80e+01	1.80e+01	1.98e+01	2.00e+01
5	2TMD	2TMD	2TMD	2TMD	2TMD	2TMD
	2.11e+01	2.09e+01	1.76e+01	1.93e+01	1.84e+01	2.25e+01
10	1CGW	1CGW	2TMD	2TMD	2TMD	3AAT
	1.84e+01	2.28e+01	2.02e+01	2.45e+01	1.86e+01	2.62e+01
30	1BGL	1BGL	1BGL	1BGL	1CGW	3AAT
	6.44e-01	7.60e-01	1.89e+00	3.17e+00	1.24e+01	1.82e+01
50	1COM	1COM	1BGL	1BGL	1CGY	1CGY
	1.73e+00	4.65e+00	7.82e-01	5.00e-01	1.42e+00	1.90e+01
70	2MTA	1DSN	1GPM	1GPM	1MMO	1TKB
	2.98e+00	6.74e+00	4.59e+00	3.08e+00	4.10e+00	1.37e+01
90	1DPG	1BSR	1GPY	1PRJ	1TKC	2TMD
	2.97e+00	1.13e+01	2.67e+00	3.14e+00	3.27e+01	5.46e+01
95	1DPG	1PYG	9GPB	1PRJ	1CYG	1BGL
	2.25e+00	4.80e+00	1.08e+00	3.54e+00	5.16e+01	6.06e+01
100	1DPG	1PYG	1PYG	1PYG	1BGL	1BGL
	3.08e+00	2.09e+00	1.32e+00	1.83e+01	4.39e+01	5.19e+01

Table 5.5: *PDB identifier and expected frequency for each search with* `t0022`.

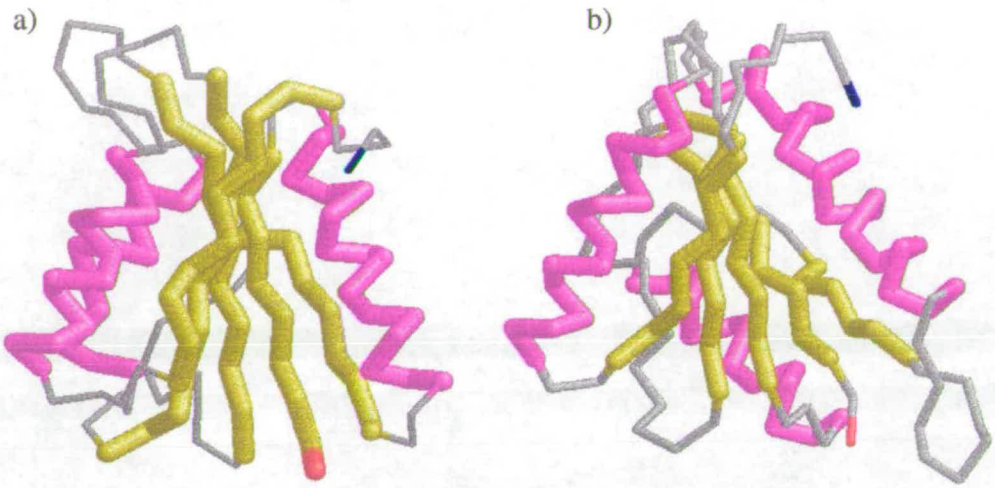


Figure 5.14: *Comparison of the homologous regions of the structures of* a) 2FX2 *and* b) t0022

The grid search (Figure 5.6) shows a couple of regions which look significant but on examination the hits in the bottom left corner are too short to form a credible prediction, as were the hits on 6APR and 1GOH. This leaves the alignment against 1PPF which covers almost the whole length of the target sequence and is surrounded in the grid search by independent serine proteases such as 1TRY.

The alignment of t0031 against 1PPF (Figure 5.16) shows low sequence identity of 15.66% but conservation is more marked in the predicted secondary structure elements. Gaps appear in predicted loops on the whole and the structural identity is 83.33%.

Release of the structure for t0031 revealed that 1PPF was the best structure to base a model on (Figure 5.17).

5.4 Discussion

The program sss_align successfully identified most of the identifiable targets for which predictions were submitted. It failed with t0020 and t0022.

In the cases where the prediction was successful there is a recognisable sequence similarity between the target sequence and the database entry. This sug-

```
                1                                                   50
predict_e942    MNNSKIISKV LLSLSLFTVG ASAFVIQDEL MQKNHAKAEV SAEEIKKHEE
ETA_STAAU P0933 MNNSKIISKV LLSLSLFTVG ASAFVIQDEL MQKNHAKAEV SAEEIKKHEE
A46569 A46569 e MNISKIISKV LLSLSLFTVG ASAFVIQDEL MQKNHAKAEV SAEEIKKHEE
ETB_STAAU P0933 .......... .......... .......... .......... ..........
SAU60589_1      .......... .......... .......... .......... ..........
PRSASK A26812;A .......... .......... .......... .......... ..........
S21758 S21758   .......... .......... .......... .......... ..........
SAU63529_1      .......... .......... .......... .......... ..........
EFSPREG_1 Z1229 .......... .......... .......... .......... ..........

                51                                                  100
predict_e942    KWNKYYGVNA FNLPKELFSK VDEKDRQKYP YNTIGNVFVK GQTSATGVLI
ETA_STAAU P0933 KWNKYYGVNA FNLPKELFSK VDEKDRQKYP YNTIGNVFVK GQTSATGVLI
A46569 A46569 e KWNKYYGVNA FNLPKELFSK VDEKDRQKYP YNTIGNVFVK GQTSATGVLI
ETB_STAAU P0933 .......... ....KELYTH ITDNARS..P YNSVGTVFVK GSTLATGVLI
SAU60589_1      .......... .......... .......... .......... ...SATGFVV
PRSASK A26812;A .......... .......... .......... .......... ....ASGVVV
S21758 S21758   .......... .......... .......... .......... ....ASGVVV
SAU63529_1      .......... .......... .......... .......... ....ATGFVI
EFSPREG_1 Z1229 .......... .......... .......... .......... .....TGFVV

                101                                                 150
predict_e942    GKNTVLTNRH IAKFANGDPS KVSFRPSINT DDNGNTETPY GEYEVKEILQ
ETA_STAAU P0933 GKNTVLTNRH IAKFANGDPS KVSFRPSINT DDNGNTETPY GEYEVKEILQ
A46569 A46569 e GKNTVLTNRH IAKFANGDPS KVSFRPSINT DDNGNTETPY GEYEVKEILQ
ETB_STAAU P0933 GKNTIVTNYH VAREAAKNPS NIIFTPAQNR DAEKNepTPY GKFEAEEIKE
SAU60589_1      GKNTILTNKH VSkyKVGDRI T....AHPNS DKGNG..... GIYSIKKIIN
PRSASK A26812;A GKDTLLTNKH VVDATHGDPH AlaFPSAINQ DNYPN..... GGFTAEQITK
S21758 S21758   GKDTLLTNKH VVDATHGDPH AlaFPSAINQ DNYPN..... GGFTAEQITK
SAU63529_1      GKNTIITNKH VSkyKVGDRI T....AHPNG DKGNG..... GIYKIKSISD
EFSPREG_1 Z1229 GTNTIVTNNH VAESFKNAKV ...LNPNAKD DAwdGSATPF GKFKVIDVA.

                151                                                 200
predict_e942    EPFGAGVDLA LIRLKPDQNG VSLGDKISPA KIGTSNDLKD GDKLELIGYP
ETA_STAAU P0933 EPFGAGVDLA LIRLKPDQNG VSLGDKISPA KIGTSNDLKD GDKLELIGYP
A46569 A46569 e EPFGAGVDLA LIRLKPDQNG VSLGDKISPA KIGTSNDLKD GDKLELIGYP
ETB_STAAU P0933 SPYGQGLDLA IIKLKPNEKG ESAGDLIQPA NIPDHIDIQK GDKYSLLGYP
SAU60589_1      ..YPGKEDVS VIQVEERakG FNFNDNVTPF KYAA..GAKA GERIKVIGYP
PRSASK A26812;A ..YSGEGDLA IVKFSPNEQN KHIGEVVKPA TMSNNAETQV NQNITVTGYP
S21758 S21758   ..YSGEGDLA IVKFSPNEQN KHIGEVVKPA TMSNNAETQV NQNITVTGYP
SAU63529_1      ..YPGDEDIS VMNIEEQakG FNFNENVQAF NFAK..DAKV DDKIKVIGYP
EFSPREG_1 Z1229 ..FSPNADIA VVTVGKQndG PELGEILTPF VLKKFES..S DTHVTISGYP

                201                                                 250
predict_e942    FDHKVNQMHR SEIELTTLSR GLRYYGFTVP GNSGSGIFNS NGELVGIHSS
ETA_STAAU P0933 FDHKVNQMHR SEIELTTLSR GLRYYGFTVP GNSGSGIFNS NGELVGIHSS
A46569 A46569 e FDHKVNQMHR SEIELTTLSR GLRYYGFTVP GNSGSGIFNS NGELVGIHSR
ETB_STAAU P0933 YNYSAYSLYQ SQIEMFNDS. ..QYFGYTEV GNSGSGIFNL KGELIGIHSG
SAU60589_1      HPYKNkvLYE STGPVMSVEg sIVYSAHTES GNSGSPVLNS NNELVGIHFA
PRSASK A26812;A GDKPVATMWE SKGKITYLKg aMQYDLSTTG GNSGSPVFNE KNEVIGIHWG
S21758 S21758   GDKPVATMWE SKGKITYLKg aMQYDLSTTG GNSGSPVFNE KNEVIGIHWG
SAU63529_1      LPAQNsqFES TGTIKRIKDN ILNFDAYIEP GNSGSPVLNS NNEVIGV...
EFSPREG_1 Z1229 GEKNHTQWSH ENDLFTseNP LLFYDIDTTG GQSGSPIYNA QFEVVGVHSN

                251                         280
predict_e942    KVSHLDREHQ INYGVGIGNY VKRIINEKNE
ETA_STAAU P0933 KVSHLDREHQ INYGVGIGNY VKRIINEKNE
A46569 A46569 e KVSHIDREHQ INYGVGIGNY VKRIINEKNE
ETB_STAAU P0933 K......... .......... ..........
SAU60589_1      SDVKNDDNRN .AYGVYFTPE IKKFIAEN..
PRSASK A26812;A GVPN...... .......... ..........
S21758 S21758   GVPN...... .......... ..........
SAU63529_1      .......... .......... ..........
EFSPREG_1 Z1229 .......... .......... ..........
```

Figure 5.15: *Multiple alignment of* t0031 *and its close homologues as used as input for* PHD *secondary structure prediction.*

Q/PAMS	40	80	110	150	250	350
0	1ASP	1ASP	1ASP	1ASP	1ASP	1ASP
	1.46e+01	1.40e+01	2.02e+01	1.46e+01	3.62e+01	1.30e+01
5	1ASP	1ASP	1ASP	1ASP	1ASP	1ASP
	1.89e+01	1.76e+01	2.04e+01	1.72e+01	2.90e+01	1.52e+01
10	1ASQ	1ASQ	1ASQ	1ASQ	1ASQ	1ASQ
	4.08e+01	3.24e+01	2.26e+01	2.37e+01	2.19e+01	1.86e+01
30	1GLK	1ASY	1ASY	1ASY	1AOZ	1PPF
	1.04e+02	5.62e+01	4.74e+01	3.68e+01	1.78e+01	8.83e+00
50	1RNR	1RNR	1ASY	1ASY	1AOZ	1PPF
	3.33e+01	2.95e+01	2.36e+01	5.17e+00	3.80e+00	3.01e-01
70	1NHK	6APR	6APR	6APR	1TRY	1PPF
	1.32e+01	1.21e+01	1.86e+00	9.53e-01	1.78e+00	7.80e-01
90	1NHK	1NHK	1GOH	1GOH	1LMW	1CAF
	7.27e-02	5.58e-01	1.22e+00	1.34e+00	2.73e+01	7.03e+01
95	1STT	1STT	1GOH	1GOH	1OVA	1CAF
	5.04e-02	7.23e-01	1.07e+00	2.75e+00	7.86e+01	1.08e+02
100	1STT	1STT	1GOH	4BLM	1OVA	1CAF
	6.20e-03	4.11e-01	2.44e+00	9.69e+00	1.19e+02	1.44e+02

Table 5.6: *PDB identifier and expected frequency for each search with* `t0031`.

```
                                                    Shane S. Sturrock
                                                          1997
                                    ALIGN 2.6
                                                  Biocomputing Research Unit
                                                  University of Edinburgh
                                                      Scotland, UK
 Sequence & Secondary Structures

PARAMETERS        t0031 - 280 residues

      pams        250              gapopen     8            gapextend   1
      rng    71   280              qval        60           phd_score VAR

STATISTICS        Mean 24.371;  Variance 110.516;  scale 0.221


SUMMARY

    1     85  1.71e-01 HYDROLASE SERINE PROTEINASE                    10-APR-89   1PPF
    2     77  6.65e-01 HYDROLASE  SERINE PROTEINASE                   07-MAR-94   1TRY
    3     66  4.02e+00 COMPLEX  SERINE PROTEASE/INHIBITOR             01-MAY-95   1BMA
    4     64  5.53e+00 HYDROLASE  SERINE PROTEINASE                   13-APR-88   1SGT


ALIGNMENTS

Result 1 - Score 85  Pred No. 1.71e-01
CHAIN E
HEADER     HYDROLASE SERINE PROTEINASE                  10-APR-89   1PPF
COMPND     HUMAN LEUKOCYTE ELASTASE  HLE   NEUTROPHIL ELASTASE  HNE
SOURCE     ELASTASE: HUMAN  HOMO SAPIENS  LEUKOCYTE; INHIBITOR: TURKEY
AUTHOR     W.BODE,A-*Z.WEI
DBLEN      218

  SEQ IDENTITY 15.66%;  SEQ CONSERVATION 48.99%;  STR IDENTITY 83.33%;

  Matches    31;  Conservative   66;  Mismatches 101;  Indels  27;  Gaps   7;

Db     17 VGGRRARPHAWPFMVSLQLRGGHFCGATLIAPNFVMSAAHCVANVNVRAVRVVL----GA 70
          |:.:  : .::. : :: ::|    .:.||: | |::. | .  .|  : :| :    .:
Qy     71 VDEKDRQKYPYNTIGNVFVKGQTSATGVLIGKNTVLTNRHIAKFANGDPSKVSFRPSINT 130

Db     71 HNLSRREPTRQVFAVQRIFENGYDPVNLLNDIVILQL---NGSATINANVQVAQLPAQGR 128
          :: :. |..   :.|: |:::.:::     |:.:::|   :.:.::..::   |::.: :.
Qy    131 DDNGNTETPYGEYEVKEILIEPPHAG---VDLALIRLKPDDNGVSLGDKIIPACIG-EN 186

Db    129 RLGNGVQCLAMGWGLLGRNRGIAEVLQELNVTVVTSLCRRSNVCTLVRGRQAGVCFGDSG 196
          | :| :   :|:.:  : . : .  ||..     |.|  . ..:. . ::|  : :|.
Qy    187 DLKDGDKLELIGYPFDHKVQCMHPSEIELTT-----LSRGLRYYGFMVPGNSGSGIFNSN 241

Db    197 SPLVCNGLIHGIASFVRGGCASGLYPD-AFAPVAQFVNWIDSIIQ 243
          :.||      :::: .| | .: .:.  . : |::..||:
Qy    242 GELVD----------IHSSKVSHLLDREHQINYGVGIGNYVKRIN 276
```

Figure 5.16: *Alignment of* 1PPF (Db) *and* t0031 (Qy).

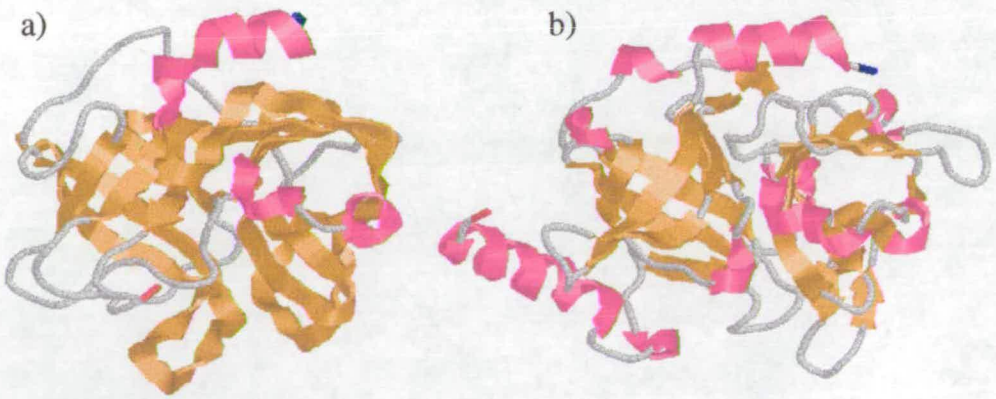


Figure 5.17: *Comparison of the structures of* a) 1PPF *and* b) t0031

gests structural similarity by evolutionary divergence from a common ancestor. However, t0014 is only aligned against part of a larger protein and, being a TIM barrel, it is likely that in this case the similarity between the structures is analogous rather than homologous. The alignment against 1BGL does have the lowest average RMS deviation between the models and the real structure which suggests that the sequence similarity does have a bearing on the similarity of the two aligned fragments of structure even though it is unlikely that they evolved from a common ancestor.

The two failures highlight the fact that sss_align is still a sequence similarity method. Target t0020, despite the high level of agreement between the secondary structure prediction and the real secondary structure, does not have enough sequence similarity to produce a correct alignment against the structure most closely similar according to structural alignment.

Because the expected frequency calculation is based on a Poisson distribution fit its behaviour in the regions of the grid search dominated by secondary structure is undefined. It is unrealistic to expect the searches with these parameter settings to be very sensitive because of the small secondary structure 'alphabet'. The best results are obtained where some sequence similarity exists which means that the correct hit is already somewhere near the top of the hit list for a sequence search only. The added value of using sss_align is that the user can gain some noise reduction based on the secondary structure prediction which can pull the real hit up out of the noise and place sequences which may have similar levels of identity down into the noise distribution if their secondary structures do not match the prediction.

Expected frequency also exaggerates the significance of short, highly matched hits with low PAM table settings and high QVAL.

Chapter 6

Conclusion

Predicting tertiary structure for a protein must be approached differently in every case. It is necessary to use a variety of tools to be successful.

Automatic mutation of structures is useful to be able to show where residues would be placed on a template structure but to take the modelling to the all atom stage is difficult and not really justified based on the likelihood that a model based on anything other than a close homologue is going to be incorrect.

The program `sss_align` has shown that it is possible to align very divergent sequences with the addition of secondary structure information. Where this approach differs from previous methods is that it can be used to align distant sequences which only have predicted secondary structures. In the case of the S-subunit alignment (Figure 3.9) the program has produced an alignment which is being verified by laboratory experiments.

In alignments against known structures, where `sss_align` has shown a result to be significant and the region of similarity covers a number of secondary structure subunits it has been shown that the sequence structure alignment has a high accuracy. It is possible that a structure which has high similarity to a new sequence exists in the database but is a homomorph. In this case fold threading would be expected to perform better than the sequence based `sss_align`.

The method for calculating significance was suitable for sequence based analysis but has proven less suited to this method. Addition of too much secondary struc-

ture information results in a distribution which does not fit the Poisson model giving unexpected statistics. This means that for the most successful results a search should still have scoring based more on sequence.

The program is a useful tool for the detection of remote sequence similarity to known structures due to its speed and simplicity.

Bibliography

[1] Abadjieva A., Patel J., Webb M., Zinkevich V., and Firman K. A deletion mutant of the type IC restriction endonuclease *Eco*R124I expressing a novel DNA specificity. *Nucleic Acids Research*, 21:4435–4443, 1993.

[2] Acharya K.R., Shapiro R. Allen S.C., Riordan J.F., and Vallee B.L. Crystal-structure of human angiogenin reveals the structural basis for its functional divergence from ribonuclease. *Proceedings of the National Academy of Science, USA*, 91:2915–2919, 1994.

[3] Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[4] Argos P. Evidence for a repeating domain in type I restriction enzymes. *EMBO Journal*, 4:1351–1355, 1985.

[5] Barcus V.A. and Murray N.E. Barriers to recombination: Restriction. In Baumberg S., Young J.P.W., Saunders S.R., and Saunders E.M.H., editors, *Population Genetics of Bacteria*, pages 31–58. Society for General Microbiology Symposium, 1995.

[6] Bickle T.A. and Kruger D.H. Biology of DNA restriction. *Microbiology Review*, 57:434–450, 1993.

[7] Bränden C. and Tooze J. *Introduction to protein structure.* Garland Publishing, Inc., USA, 1991.

[8] Burckhardt J., Weisemann J., Hamilton D.L., and Yuan R. Complexes formed between the restriction endonuclease *Eco*K and heteroduplex DNA. *Journal of Molecular Biology*, 153:425–440, 1981.

[9] Chen A., Powell L.M., Dryden D.T.F., Murray N.E., and Brown T. Tyrosine 27 of the specificity polypeptide of *Eco*KI can be UV crosslinked to a bromodeoxyuridine-substituted DNA target sequence. *Nucleic Acids Research*, 23:1177–1183, 1995.

[10] Cheng X., Kumar S., Posfai J., Pflugrath J.W., and Roberts R. Crystal structure of the *Hha*I DNA methyltransferase complexed with S-adenosyl-L-methionine. *Cell*, 74:299–307, 1993.

[11] Cheng X. and Blumenthal R.M. Finding a basis for flipping bases. *Current Opinion in Biology*, 4:639–645, 1996.

[12] Chothia C. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[13] Chou P.Y. and Fasman G.D. Conformational parameters for amino acids in helical, β strand and random coil regions calculated from proteins. *Biochemistry*, 13:211–222, 1974a.

[14] Chou P.Y. and Fasman G.D. Predicton of protein conformation. *Biochemistry*, 13:222–245, 1974b.

[15] Collins J.F.C. and Coulson A.F.W. Molecular sequence comparison and alignment. In Bishop M.J. and Rawlings C.J., editors, *Nucleic acid and protein sequence analysis: A practical approach*, chapter 13, pages 323–358. IRL Press, 1987.

[16] Collins J.F.C. and Coulson A.F.W. Significance of protein sequence similarities. *Methods in Enzymology*, 183:474–487, 1990.

[17] Conti E., Franks N.P., and Brick P. Crystal-structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure*, 4:287–298, 1996.

[18] Cooper L.P. and Dryden D.T.F. The domains of a type I DNA methyltransferase: interactions and role in recognition of DNA methylation. *Journal of Molecular Biology*, 236:1011–1021, 1994.

[19] Coulson A.F.W., Collins J.F., and Lyall A. Protein and nucleic acid sequence database searching: A suitable case for parallel processing. *Computer Journal*, 30:420–423, 1987.

[20] Cowan G.M., Gann A.A.F., and Murray N.E. Conservation of complex DNA recognition domains between families of restriction enzymes. *Cell*, 56:103–109, 1989.

[21] Creighton T.E. The energetic ups and downs of protein folding. *Nature Structural Biology*, 1:135–138, 1994.

[22] Dayhoff M.O., Schwartz R.M., and Orcutt B.C. *Atlas of Protein Sequence and Structure*, volume 5, suppl. 3, pages 345–352. National Biomedical Research Foundation, Washington, DC., 1978.

[23] Dittmann J., Wenger R.M., Kleinkauf H., and Lawen A. Mechanism of cyclosporine-a biosynthesis — evidence for synthesis via a single linear undecapeptide precursor. *Journal of Biological Chemistry*, 269:2841–2846, 1994.

[24] Dryden D.T.F., Sturrock S.S., and Winter M. Structural modelling of a type I DNA methyltransferase. *Nature Structural Biology*, 2:632–635, 1995.

[25] Dryden D.T.F., Cooper L.P., Thorpe P.H., and Byron O. The in vitro assembly of the *Eco*KI type I DNA restriction/modification enzyme and its in vivo implications. *Biochemistry*, 36:1065–1076, 1997.

[26] Dunbrack RL. and Karplus M. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *Journal of Molecular Biology*, 230:543–574, 1993.

[27] Dybvig K. and Yu H. Regulation of a restriction and modification system via DNA inversion in *mycoplasma pulmonis*. *Molecular Microbiology*, 12:547–560, 1994.

[28] Eisenberg D., Bowie J.U., Lüthy R., and Choe S. 3-dimensional profiles for analyzing protein-sequence structure relationships. *Faraday Discussions*, pages 25–34, 1992.

[29] Eisenmenger F., Argos P., and Abagyan R. A method to configure protein side-chains from the main-chain trace in homology modelling. *Journal of Molecular Biology*, 213:849–860, 1993.

[30] Fenton W.A. and Horwich A.L. GroEL-mediated protein folding. *Protein Science*, 6:743–760, 1997.

[31] Gann A.A.F., Campbell A.J.B., Collins J.F., Coulson A.F.W., and Murray N.E. Reassortment of dna recognition domains and the evolution of new specificities. *Molecular Microbiology*, 1:13–22, 1987.

[32] Garnier J., Osguthorpe D.J., and Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120:97–120, 1978.

[33] Gotoh O. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[34] Gribskov M., Lüthy R., and Eisenberg D. Profile analysis. *Methods in Enzymology*, 188:146–159, 1990.

[35] Gribskov M., McLachlan A.D., and Eisenberg D. Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Science, USA*, 84:4355–4358, 1987.

[36] Gubler M., Braguglia D., Meyer J., Piekarowicz A., and T.A. Bickle. Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction-modification enzymes. *EMBO Journal*, 11:233–240, 1992.

[37] Henikoff S. and Henikoff J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science, USA*, 89:10915–10919, 1992.

[38] Holm L. and Sander C. Fast and simple monte carlo algorithm for side chain optimization in proteins: Application to model building by homology. *Proteins: Structure, Function and Genetics*, 14:213–223, 1992.

[39] Jones D., Taylor W.R., and Thornton J.M. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[40] Jones D. and Thornton J.M. Protein fold recognition. *Journal of Computer-Aided Molecular Design*, 7:439–456, 1993.

[41] Kabsch W. and Sander C. A dictionary of protein secondary structure. *Bioplymers*, 22:2577–2637, 1983.

[42] Kannan P., Cowan G.M., Daniel A.S., Gann A.A.F., and Murray N.E. Conservation of organisation in the specificity polypeptides of two families of type I restriction enzymes. *Journal of Molecular Biology*, 209:335–344, 1989.

[43] Karreman C. and de Waard A. *Agmenellum quadruplicatum* M.*Aqu*I, a novel modification methylase. *Journal of Bacteriology*, 172:266–272, 1990.

[44] Kelleher J.E., Daniel A.S., and Murray N.E. Mutations that confer de novo activity upon a maintenance methyltransferase. *Journal of Molecular Biology*, 221:431–440, 1991.

[45] King G. and Murray N.E. Restriction enzymes in cells, not eppendorfs. *Trends in Microbiology*, 2:465–469, 1994.

[46] Klimasauskas S., Kumar S., Roberts R., and Cheng X. *Hha*I methyltransferase flips its target base out of the DNA helix. *Cell*, 76:357–369, 1994.

[47] Korona R. and Levin B.R. Phage-mediated selection and the evolution and maintenance of restriction-modification. *Evolution*, 47:556–575, 1993.

[48] Kumar S., Cheng X.D., Klimasauskas S., Mi S., Posfai J., Roberts R.J., and Wilson G.G. The DNA (Cytosine-5) methyltransferases. *Nucleic Acids Research*, 22:1–10, 1994.

[49] Labahn J., Granzin J., Schluckebier G., Robinson D.P., Jack W.E., Schildkraut I., and Saenger W. Three dimensional structure of the adenine specific DNA methyltransferase M.*Taq*I in complex with the cofactor S-adenosylmethionine. *Proceedings of the National Academy of Science, USA*, 91:10957–10961, 1994.

[50] Lauster R. Evolution of type II DNA methyltransferases: A gene duplication model. *Journal of Molecular Biology*, 206:313–321, 1989.

[51] Lee K.F., Kam K.M., and Shaw P.C. A bacterial methyltransferase M.*Eco*HK31I requires two proteins for in vitro methylation. *Nucleic Acids Research*, 223:103–108, 1995.

[52] Levin B.R. Frequency dependent selection in bacterial populations. *Philosophical Transactions of the Royal Society of London Series B*, 319:459–472, 1988.

[53] Lüthy R., Bowie J.U., and Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature*, 356:83–85, 1992.

[54] Malone T., Blumenthal R.M., and Cheng X. Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyl-transferases, and suggests a catalytic mechanism for these enzymes. *Journal of Molecular Biology*, 253:618–632, 1995.

[55] Meister J., MacWilliams M., Hubner P., Jutte H., Skrzypek E., Piekarowicz A., and Bickle T.A. Macroevolution by transposition: drastic modification of DNA recognition by a type I restriction enzyme following Tn5 transposition. *EMBO Journal*, 12:4585–4591, 1993.

[56] Murray N.E., Daniel A.S., Cowan G.M., and Sharp P.M. Conservation of motifs within the unusually variable polypeptide sequences of type I restriction and modification enzymes. *Molecular Microbiology*, 9:133–143, 1993.

[57] Needleman S.B. and Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[58] Noyer-Weidner M., Walter J., Terschüren P.A., Chai S., and Trautner T.A. A new monospecific DNA (Cytosine-C5) methyltransferase with pronounced amino-acid-sequence similarity to a family of Adenine-N6-DNA-Methyltransferases. *Nucleic Acids Research*, 22:4066–4072, 1994.

[59] Pearson W.R. and Lipman D.J. Improved toos for biological sequence comparison. *Proceedings of the National Academy of Science, USA*, 85:2444–2448, 1988.

[60] Ponder J.W. and Richards F.M. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193:775–791, 1987.

[61] Powell L.M., Dryden D.T.F., Willcock D.F., Pain R.H., and Murray N.E. DNA recognition by the *Eco*K methyltransferase. *Journal of Molecular Biology*, 234:60–71, 1993.

[62] Powell L.W. and Murray N.E. S-Adenosyl Methionine alters the DNA contacts of the *Eco*KI methyltransferase. *Nucleic Acids Research*, 23:967–974, 1995.

[63] Price C., Lingner J., Bickle T.A., Firman K., and Glover S.W. Basis for changes in DNA recognition by the *Eco*R124 and *Eco*R124/3 type I DNA restriction and modification enzymes. *Journal of Molecular Biology*, 205:115–125, 1989.

[64] Roberts R.J. On base flipping. *Cell*, 82:9–12, 1995.

[65] Rossman M.G., Liljas A., Bränden C.I., and Banaszak L.J. *The Enzymes*, volume XI. Boyer P.D. (ed), Academic Press, 3rd edition, 1975.

[66] Rost B. TOPITS: Threading 1-dimensional predictions into 3-dimensional structures. In Rawlings C., Clark D., Altman R., Hunter L., Lengauer T., and Wodak S., editors, *The third international conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 314–321. Menlo Park, CA: AAAI Press, 1995.

[67] Rost B. and Sander C. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–599, 1993.

[68] Russell R.B., Copley R.R., and Barton G.J. Protein fold recognition by mapping predicted secondary structures. *Journal of Molecular Biology*, 259:349–365, 1996.

[69] Sander C. and Schneider R. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[70] Sharp P.M., Kelleher J.E., Daniel A.S., Cowan G.M., and Murray N.E. Roles of selection and recombination in the evolution of type I restriction-modification systems in enterobacteria. *Proceedings of the National Academy of Science, USA*, 89:9836–9840, 1992.

[71] Sippl M.J. Calculation of conformational ensembles from potentials of mean force. *Journal of Molecular Biology*, 213:859–883, 1990.

[72] Smith T.F. and Waterman M.S. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[73] Sosnik T.R., Mayne L., Hiller R., and Englander S.W. The barriers in protein folding. *Nature Structural Biology*, 1, No. 3:149–156, 1994.

[74] Taylor I.A., Davis K.G., Watts D., and Kneale G.G. DNA-binding induces a major structural transition in a type I methyltransferase. *EMBO Journal*, 13:5772–5778, 1994.

[75] Taylor I.A., Webb M., and Kneale G.G. Surface labelling of the type I methyltransferase M.*Eco*R124I reveals lysine residues critical for DNA binding. *Journal of Molecular Biology*, 258:62–73, 1996.

[76] Taylor W.R. and Orengo C.A. Protein structure alignment. *Journal of Molecular Biology*, 208:1–22, 1989.

[77] Thorpe P.H., Ternent D., and Murray N.E. The specificity of *Sty*SKI, a type I restriction enzyme, implies a structure with rotational symmetry. *Nucleic Acids Research*, 25:1694–1700, 1997.

[78] Titheradge A.J.B., Ternent D., and Murray N.E. A third family of allelic *hsd* genes in *salmonella enterica* – sequence comparisons with related proteins identify conserved regions implicated in restriction of DNA. *Molecular Microbiology*, 22:437–447, 1996.

[79] vanVlijmen H.W.T. and Karplus M. PDB-based protein loop prediction: Parameters for selection and methods for optimization. *Journal of Molecular Biology*, 267:975–1001, 1997.

[80] Waterman M.S., editor. *Mathematical methods for DNA sequences.* CRC Pres, Boca Raton, Florida, USA, 1988.

[81] Webb M., Taylor I.A., Firman K., and Kneale G.G. Probing the domain structure of the type IC DNA methyltransferase M.*Eco*R124I by limited proteolysis. *Journal of Molecular Biology*, 250:181–190, 1995.

[82] Weber GG, Schorgendorfer K., Scheiderscherzer E., and Leitner E. The peptide synthetase catalyzing cyclosporine production in *tolypocladium-niveum* is encoded by a giant 45.8-kilobase open reading frame. *Current Genetics*, 26:120–125, 1994.

[83] Willcock D.F., Dryden D.T.F., and Murray N.E. A mutational analysis of the 2 motifs common to adenine methyltransferases. *EMBO Journal*, 13:3902–3908, 1994.

[84] Wilmanns M. and Eisenberg D. Three-dimensional profiles from residue-pair preferences: Identification of sequences from β/α-barrel fold. *Proceedings of the National Academy of Science, USA*, 90:1379–1383, 1993.

[85] Wilson G.G. Amino-acid-sequence arrangements of DNA-methyltransferases. *Methods in Enzymology*, 216:259–279, 1992.

[86] Wilson G.G. and Murray N.E. Restriction and modification systems. *Annual Review of Genetics*, 138:283–288, 1991.

Sequence and Secondary Structure Alignment with 'sss_align': Shane S. Sturrock, 1997.

This is a guide to using and getting the best out of `sss_align`. It is rather rough and ready but should be of some use. I am making this code available for people to try it and give me feedback because of a number of requests for copies, we are working on a paper fully describing the program for reference purposes. The present installation method is not ideal but it should get you going, I hope you find something of interest in the database. The tar file can be downloaded from http://www.icmb.ed.ac.uk/sss_align

1. Introduction

The program `sss_align` is an implementation of the Smith/Waterman best local similarity algorithm which, in addition to the traditional Dayhoff PAM scoring scheme, has the ability to use secondary structure prediction in order to align sequences which have identities as low as 15% and still identify those sequences as significant. Previously, this level of sensitivity was the preserve of the 'fold-threading' algorithms.

The program can run in a number of modes:

Standard Smith/Waterman search
: Given a database in either FASTA, SWISS/EMBL, PHD or SSSDB format (such as `sssdb` and `hsssdb`) the program will do a normal sequence search using user selectable PAMs and gap penalties if the query or database is in simple FASTA or SWISS/EMBL format.

Secondary structure search
: The query and database will both contain secondary structure information, predicted or real, database in `sssdb` or `PHD` format.

Scan search
: A very fast pre-search filter can be applied which compares sequences composition before doing a Smith/Waterman alignment so it only bothers to align sequences which have similar length and residue compositions. This can be useful when searching large databases, e.g. SwissProt, for sequences which have a reasonable level of sequence identity. This search is not as sensitive as doing a full exhaustive search so should not be used exclusively.

The program is compiled to run on Silicon Graphics workstations running IRIX 6.2, and will run on Challenge machines although only on a single processor. The database provided is made up from non-redundant sequences of PDB taken from the DSSP database, plus non-redundant sequences from HSSP giving a total of over 15200 different sequences with secondary structure information. This database was generated automatically by inverting DSSP+HSSP and removing all repeats or close

homologues (>90% ID).

2. Installation

You should have downloaded a file called `sss_align.tar.gz`. This should be uncompressed with `gunzip` and then untarred with `tar xvf sss_align.tar`.

`sss_align` and its databases should be installed in `/usr/local/bin` so that it will be easily visible to all users. The package consists of the pre-compiled `sss_align` executable, two databases (`sssdb` and `hsssdb` derived from DSSP and HSSP databases respectively), and the source code for `MPmail` which is provided for generating suitable `PHD` prediction source files. This last program should be installed on the machine each user sends and receives e-mail with and can be compiled by simply typing `cc -o MPmail MPmail.c`. It takes the filename of an `MPsrch` output file as its only parameter.

3. Up and Running

Once installed the program can be run simply by typing `sss_align` at the command prompt. To check that the program is installed correctly a test file is included in the package - `t4.phd`. Copy this file to your home directory and type the following:

```
sss_align -infile t4.phd -pams 150 -gapopen 10 -gapextend 4 -qval 90
-align 1 -summary 5
```

This will run showing you the progress of the execution (`sssdb` contains just over 2000 sequences at present) and when finished you will have a new file `t4.phd.out` in the same directory.

```
more t4.phd.out
```

```
        _/_/_/      _/_/_/      _/_/_/
     _/_/  _/_/  _/_/  _/_/  _/_/  _/_/                    Shane S.  Sturrock
    _/_/        _/_/        _/_/                               1997
     _/_/_/      _/_/_/      _/_/_/      ALIGN 2.6
        _/_/        _/_/        _/_/                  Biocomputing Research Unit
  _/_/  _/_/  _/_/  _/_/  _/_/  _/_/                  University of Edinburgh
   _/_/_/      _/_/_/      _/_/_/                         Scotland, UK
 Sequence & Secondary Structures

PARAMETERS      t0004.phd - 84 residues

      pams      150             gapopen    10            gapextend   4
      rng    1   84             qval       90            phd_score VAR

STATISTICS      Mean 28.115;  Variance 92.542;  scale 0.304


SUMMARY

   1     74  5.50e-01 TRANSCRIPTION REGULATION                    12-MAY-93   1CSP
   2     60  8.82e+00 PHOSPHOTRANSFERASE                          25-SEP-91   1GPR
   3     59  1.06e+01 DNA-BINDING PROTEIN                         31-MAR-95   1SSO
   4     58  1.28e+01 T LYMPHOCYTE ADHESION GLYCOPROTEIN          10-AUG-94   1HNF
   5     57  1.54e+01 HYDROLASE (UREA AMIDO)                      20-JUN-95   1KRA
```

```
ALIGNMENTS

Result 1 - Score 74  Pred No. 5.50e-01
CHAIN -
HEADER    TRANSCRIPTION REGULATION                  12-MAY-93   1CSP
COMPND    MAJOR COLD SHOCK PROTEIN (CSPB)
SOURCE    ORGANISM: BACILLUS SUBTILIS;
AUTHOR    H.SCHINDELIN,U.HEINEMANN
DBLEN     67

  SEQ IDENTITY 34.43%;  SEQ CONSERVATION 55.74%;  STR IDENTITY 81.97%;

  Matches    21;  Conservative   13;  Mismatches  27;  Indels   8;  Gaps   4;

           EEEEEEEttttteE EEE tt   EEEE                  tt EEEEEEEEEttee
Db       4 GKVKWFNSEKGFG-FIEVEGQDDVFVHFSAIQG---EGFKT-LEEGQAVSFEIVEGNRGP 58
           |||.:: .   || |:.: |  : :|||:| | .   |  .  |: ||.|:  ::| :|
Qy      11 GKVTRIVD---FGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQG 67
           EEEEEEE     EeEEEe    eEEEEeee           eeee  EEEEEEEEE

           EEEEEEE
Db      59 QAANVTKEA 67
           :    .|||
Qy      68 RIRLSIKEA 76
           EEEEEEE
```

This search shows a typical successful search, the top hit shows significant sequence and structure similarity and as it turns out the prediction of the structure for this sequence is correct. Of course, `sss_align` can detect similarities between sequences which have diverged far more than this example. But at least you know it works now!

4. Parameters

In this section I will describe the main parameters available and how they can affect your search.

`-infile` *file_name*

This is the only compulsory parameter and it specifies a UNIX filename. The format of this file will be the returned output of PHD, FASTA format, SWISS/EMBL or SSSDB. You could search the database using just this one parameter but the chances are that you would not find a significant hit. For example, try `sss_align -infile t4.phd` and you will find that the output file does not contain any significant hits. The top hit is still the same but the alignment is shorter which is why the signal didn't make it out of the noise. You could try the `-plot` parameter to write an histogram which will give you a better idea of what is happening.

`-pams` *1-500*

This parameter specifies the Dayhoff PAM table to use in the range 1-500, with 1 being the most strict scoring and 500 the most relaxed. Since every sequence family has diverged by different amounts no one PAM table is ideal. Typically a strict PAM table will result in hits that are short and highly conserved, the more distant PAM tables will allow the alignment to get longer but will increase the noise from false positives so you will have to try a number of tables to see what works best.

`-gapopen` *1-100*

Gapopen is the cost of opening a single gap, its optimal value varies with which PAM table is in

use so you can let the program choose one for you at first.

`-gapextend` *0-gapopen*

Gapextend is the cost of extending a gap once it is open, this is referred to as an affine gap where the cost to open the gap is typically much higher than the cost of extending one. Again, the program can be left to select a suitable extension penalty which the user can fine tune later.

`-qval` *0-100*

All the previous parameters are typical of sequence searching programs, this one is new. Each `sss_align` search usually has two scoring schemes running, the normal PAM score and a log odds table derived from the reliabilities of the predicted secondary structures in a `PHD` prediction. The qval allows the user to define the percentage of each table the search combines to form the actual numbers used by the Smith/Waterman alignment. 100 means all sequence, 0 is all secondary structure. The best results will be obtained about 30-70 but try it and see.

`-summary` *0-max*

Specifies the number of hits in the output summary list.

`-align` *0-max*

Specifies the number of sequences for which an alignment will be produced, if a number bigger than the number of summaries is requested the summary table will cover the same range.

`-dbname` *file_name*

This allows you to specify a particular database, the default is `/usr/local/bin/sssdb`, an alternative is which contains the non-redundant version of HSSP. This database is contains over 15000 sequences so searches will take a bit longer but you have a better chance of finding a close homologue. You can also use your own databases in FASTA format for normal sequence searches.

You should be able to complete most of your searches using the above set of parameters, try them with the test file to see why the parameters I gave for your first run were best. To see what others are available you can type `sss_align -help` and it will give you a list with ranges and brief descriptions.

5. Making your own query

While any `PHD` prediction will do just fine, and if you have a prefered way of generating these then more power to you, I suggest you try using the following method. The program `MPsrch` is a very fast implementation of the Smith/Waterman algorithm and for anyone who is using `sss_align` you will find it very familiar indeed since I worked on this program too in its pre-commercial days. Thus I heartily recommend it as the first stop for generating a good list of related sequences to pass to the `PHD` server rather than letting it do a search of its own. The advantage of using `MPsrch` is that you have full control over the parameters just as with `sss_align` and it can search very large protein databanks extremely quickly. A very good `MPsrch` service can be found at this Japanese site which has a 16384 processor MP2 MasPar system - http://www.dna.affrc.go.jp/htdocs/MPsrch/index.html

This is also a great site for every day database searching rather than relying on heuristic methods such as BLAST and FASTA since it is quicker and more sensitive than either. If you use this service you can feed the returned searches straight into the program `MPmail` which is provided with `sss_align` and it will automatically generate a file which can simply be e-mailed to the `PHD` server by typing:

```
mail predictprotein@embl-heidelberg.de < file_name
```

Shortly you should receive an e-mail message containing the prediction which can be saved as a file and used as the input to `sss_align`.

MPmail takes care of only sending sequences which MPsrch shows as being significant and also deletes any fragments which don't tend to add information to a multiple alignment. As a result you should get back a very good secondary structure prediction.

6. Algorithm

As already stated, this program uses the standard Smith/Waterman best local similarity algorithm and to find out how that works you can look up the reference. The implementation does not include any cheats to force gaps in loops for instance, it is a pure Smith/Waterman.

The difference between sss_align and a normal implementation like MPsrch is the added information from secondary structure prediction. While the alignment still requires that the sequences be divergently evolved it becomes possible to detect similarities which are much more distant than previously possible using a simple one-dimensional search because the reliable secondary structure predictions produced by PHD allow sequences which would normally have been in the twilight zone of a sequence search to get a boost from similar secondary structure and thus climb to significance while others which don't have similar secondary structures become less significant.

6.1 Prediction scoring

In order to use secondary structure predictions successfully to build a log odds scoring scheme it was necessary to know how reliable the prediction was. This information is available with PHD predictions but it isn't enough just to know that on average the predictions are 70% reliable, for this method to really work you must know on a residue by residue basis how likely it is that a particular residue predicted to be of secondary structure type X is in fact X and how likely it is that it is Y or Z for instance. Also, it is important to be able to score residues where the secondary structures don't match and not to penalise a mismatch too much where the predicted residue does not have an high confidence. Given that you know the reliability of the secondary structure of each position in both sequences it becomes possible to construct a log odds table for matches and mismatches as an alignment is made.

6.2 Reliabilities

With sss_align searches, the database entries are all either real PDB entries or close homologues and so can be considered to have reliable secondary structures, so for each residue the reliability is 100%. The PHD prediction does not have this luxury and in this case it is necessary to retrieve the position specific reliabilities on a residue by residue basis and turn these numbers into percentages which can be entered into the algorithm (B. Rost provided the numbers to use in this translation). The advantage of this is that if PHD makes an incorrect prediction the program will not force a misalignment if the reliability of that prediction is low as it tends to be when predictions are wrong. This can be seen in the test sequence where one of the predicted strands does not exist in the homologous structure but since it has a low reliability the prediction still works:

```
                 ....,....1....,....2....,....3....,....4....,....5....,....6
        AA      |AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKV|
        PHD sec |  EEEEEEEEEEEEEEE  EEEEEE    EEEEEEEE           EEEE  EEEEEE|
        Rel sec |952665799999987313358998388732588986211201577321366285899999|
detail:
        prH sec |000000000000000000000000000000000013345441113331000000000000|
```

```
         prE sec |025776888989888553378998611135788887421000100002577512899999|
         prL sec |973222100000011345521001388753211001244344687653211487100000|
 subset: SUB sec |LL.EEEEEEEEEEE....EEEEE.LLL..EEEEEE......LLL....EE.LLEEEEEE|
```

The strand at 49-52 has a lower reliability than the other strands and in the alignment this strand has to be aligned with loop residues.

```
           EEEEEEEttteE EEE tt   EEEE                  tt EEEEEEEEEttee
Db       4 GKVKWFNSEKGFG-FIEVEGQDDVFVHFSAIQG---EGFKT-LEEGQAVSFEIVEGNRGP 58
           |||.:: .   || |:.: |  : :||:| | .   |  .  |: ||.|:  ::| :|
Qy      11 GKVTRIVD---FGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQG 67
           EEEEEEE     EeEEEe    eEEEEeee           eeee  EEEEEEEEE
```

The variable scoring ensures that such a visual misalignment does not incur a large score penalty while strongly predicted residues matching against like secondary structure provide a score bonus. This is the essence of why `sss_align` works so well.

6.3 Mixing tables

Merging the two scoring schemes in a way which was consistent with the Smith/Waterman algorithm meant that the new table had to be scaled so that its most negative score matched the PAM table so that the gap penalty suited to the PAM table would also suit this new table. This also meant that it was possible to simply take a fraction of one table and the other, add them together and produce an hybrid table which would still suit the gap penalty chosen. The fraction of each table used is determined by the `-qval` parameter.

7. Conclusion

This new program was used in a blind trial of fold recognition methods and was applied in a pure way, ie the results it said were correct were what was sent in. No biological knowledge other than sequence was used. The result was that the program identified three sequences which were possible relatives and these three did in fact prove to be correct. The program is fast and sensitive which allows exploration of the wide range of parameters in order to get the best result, this above all makes `sss_align` a very useful tool for the identification of distant homologues.

Shane Sturrock, Biocomputing Research Unit, Edinburgh, UK

Queries and comments via e-mail sss@holyrood.ed.ac.uk

A prediction of the amino acids and structures involved in DNA recognition by type I DNA restriction and modification enzymes.

Shane S. Sturrock & David T.F. Dryden

Institute of Cell & Molecular Biology

The King's Buildings

University of Edinburgh

Mayfield Road

Edinburgh

EH9 3JR

United Kingdom

Tel. +44-131-650-5378, Fax +44-131-650-8650,

Email David.Dryden@ed.ac.uk or sss@holyrood.ed.ac.uk

Abstract

The S subunits of type I DNA restriction-modification enzymes are responsible for recognising the DNA target sequence for the enzyme. They contain two domains of approximately 150 amino acids, each of which is responsible for recognising one half of the bipartite asymmetric target. In the absence of any known tertiary structure for type I enzymes or recognisable DNA recognition motifs in the highly variable amino acid sequences of the S subunits, it has previously not been possible to predict which amino acids are responsible for sequence recognition. Using a combination of sequence alignment and secondary structure prediction methods to analyse the sequences of S subunits, we predict that all of the 51 known target recognition domains (TRDs) have the same tertiary structure. Furthermore, this structure is similar to the structure of the TRD of the C5-cytosine methyltransferase, *Hha*I, which recognises its DNA target via interactions with two short polypeptide loops and a β strand. Our results predict the location of these sequence recognition structures within the TRDs of all type I S subunits.

Keywords: DNA / endonuclease / methyltransferase / specificity

Running head: Sequence specific DNA recognition by type I restriction enzymes.

Introduction

A major aim of many studies of sequence specific protein-DNA interactions has been to determine how certain sequences of amino acids can recognise, with great fidelity, a DNA target sequence. Structural analysis of protein-DNA complexes has shown how α helices, β strands and loops can be used to give sequence specificity (Harrison, 1991; Luisi, 1995; Choo & Klug, 1997).

DNA methyltransferases (mtases) of restriction/modification (R/M) systems use target recognition domains (TRD) of 50-150 amino acids to recognise their DNA target sequence (Wilke *et al*, 1988). The TRD is the major determinant in DNA target specificity with separate catalytic domains being required for enzymatic activity. The crystal structures of two monomeric type II C5-cytosine mtases, *Hha*I and *Hae*III, bound to their DNA targets show that the TRD uses a conserved structure comprising two loops and one β strand to accomplish sequence recognition (Klimasauskas *et al*, 1994; Reinisch *et al*, 1995). The amino acid sequences of TRDs of many different C5-cytosine DNA mtases have been compared. The level of sequence identity in these comparisons is very low and confined to several very short amino acid sequences corresponding to the recognition region in the two crystal structures (Cheng & Blumenthal, 1996; Lange *et al*, 1996), however, experimental support has been obtained for the involvement of this region in DNA recognition by C5-cytosine mtases other than *Hha*I and *Hae*III (Lange *et al*, 1996). The N6-adenine and C4-cytosine mtases also contain a TRD and a catalytic domain (Malone *et al*, 1995), however, no cocrystal structure of one of these enzymes with DNA has been solved. A model of DNA recognition by the *Taq*I N6-adenine mtase, whose structure is known in the absence of DNA, has been constructed (Labahn *et al*, 1994; Schluckebier *et al*, 1995).

All characterised type I R/M systems recognise N6-adenine methylation of a bipartite target sequence (see Bickle & Kruger, 1993; King & Murray, 1994; Barcus & Murray, 1995 for reviews). They are large, oligomeric, multifunctional enzymes encoded by the *hsdR*, *M* and *S* genes, combining both restriction endonuclease(R) and modification mtase(M) subunits with a DNA specificity (S) subunit. Type I R/M systems of enteric bacteria have been grouped

into four families based on subunit complementation, DNA hybridisation and antibody cross-reactivity experiments (see Barcus & Murray, 1995; Titheradge *et al*, 1996). The amino acid sequence identity is very high within a family for the R, M and parts of the S subunit outwith the TRDs (Gann *et al*, 1987; Cowan *et al*, 1989; Kannan *et al*, 1989; Sharp *et al*, 1992; Murray *et al*, 1993; Gubler *et al*, 1992). This is believed to reflect conservation of residues in the subunit interfaces and the nuclease and mtase catalytic sites.

The S subunits of type I R/M systems contain two TRDs of 150-180 amino acids (see Bickle & Kruger, 1993; King & Murray, 1994; Barcus & Murray, 1995 for reviews). Each TRD is responsible for recognising one of the two parts of the bipartite DNA target. The amount of amino acid sequence conservation between TRDs is either below approximately 20% for TRDs recognising different targets, or 40% to 90% when a target is shared dependent on whether the S subunits are in different of the same family. The remainder of the approximately 50kDa S subunit contains amino acid sequences which show a high degree of conservation between type I systems. These regions are responsible for defining the length of the non-specific DNA spacer in between the two TRD target sequences (Price *et al*, 1989) and for binding the M and R subunits (Abadjieva *et al*, 1993; Meister et al, 1993; Cooper & Dryden, 1994; Webb *et al*, 1995). Each TRD fits into the major groove to recognise the DNA, and the M subunits are arranged on either side of the S subunit allowing them to encircle the DNA and gain access to the methylation targets (Kneale, 1994; Dryden *et al*, 1995).

The TRDs of type I S subunits recognise a wide variety of 3, 4 or 5 base pair targets and it would be of interest to define which amino acids within the large and highly variable sequence of the TRDs are responsible for sequence specificity. In this paper we use amino acid sequence alignment combined with secondary structure prediction methods. The use of secondary structure predictions enhances the strength of the amino acid alignment making distant similarities more apparent. These alignments of the TRDs suggest that all have the same tertiary structure and that they are the products of divergent evolution. A comparison of the secondary structure predictions with the known structure of the TRD of the *Hha*I mtase shows a strong similarity which has allowed us to define potential DNA recognition loops for all of the type I TRDs

and to model the tertiary structure of these domains in a manner amenable to experimental verification.

Results

Normal sequence alignment methods have been applied to complete S subunit sequences in the past (Argos, 1985; Gann *et al*, 1987). These studies were hampered not only by the limited number of sequences available but also by the high degree of sequence similarity in the conserved regions of the subunits. These restricted areas of high homology almost totally obscured any sequence similarity between TRDs except when the TRDs recognised identical DNA targets whereupon the similarity was so high as to preclude any prediction of amino acids involved in sequence recognition.

The result of applying `sss_align` to the TRDs of type I S subunits is shown in figure 1. In contrast to previous analyses, we were able to observe considerable sequence similarity between TRDs even if the TRDs recognised different DNA targets. β strand 1 was used as the centre point of the multiple alignment. Using this predicted β strand to "lock" the sequences together, it is apparent that many other predicted secondary structure features, particularly loops 1 and 2 and β strand 2, then become aligned even when sequences are very distantly related. Close inspection of figure 1 suggests that some of the sequence alignments could be "improved" by the introduction of small gaps or deletions, however, in the absence of any experimental data to support such shifts, we present only the raw output of the program. Overall, we believe that these results are suggestive of a common tertiary structure for the TRDs of type I S subunits.

We compared our alignment of all 51 TRDs with the known sequence and secondary structure of the TRD of the C5-cytosine mtase *Hha*I, figure 1(Klimasauskas *et al*, 1994). To our surprise, the two loops and one β strand which are responsible for the recognition of the DNA target of *Hha*I matched very well with our alignment of type I TRDs if the β strand preceding the first recognition loop in *Hha*I, is aligned with β strand 1 of our alignment of type I TRDs. Figure

2 shows the recognition of the DNA phosphate backbone and bases by part of the TRD of *Hha*I mtase. In *Hha*I, loop 1 (Val232-Glu239) fills the major groove and positions Gln237 into the gap left by the flipped out cytosine base, β strand 2 (Arg240-Tyr242) makes important base and phosphate contacts, and Thr250-Phe259, as part of the long loop 2, makes further backbone and base contacts.

The agreement in length and composition of strand 1 is good between *Hha*I and all of the type I TRDs. Loop 1 is generally predicted to be shorter and β strand 2 longer than equivalent structures in *Hha*I. However, in the prediction for *Eco*KI-1 for example, the three amino acid long strand 2 is preceded by an extra predicted strand which may suggest that the extra length of strand 2 in many of our predictions is due to a tendency for PHD to overpredict the length of a strand or to merge two strands together. In *Hha*I, strand 2 commences with Arg240 and it is apparent that an equivalent basic amino acid, e.g. Lys92 in *Eco*KI, is present in many of the type I TRDs, though usually in the middle of the longer predicted strand 2. Arg240 in *Hha*I is involved in base recognition and perhaps suggests a similar role for these basic residues in type I S subunits. Loop 2 is 21 amino acids long in the *Hha*I-DNA cocrystal structure but in the absence of DNA, the loop is interrupted by a β strand at amino acids 250-253 (Cheng *et al*, 1993; Klimasauskas *et al*, 1994). The existence of an equivalent extra β strand in the middle of loop 2 is predicted for many of the type I TRDs. In *Hha*I mtase, loop 2 terminates with another β strand, however, many of our predictions suggest that in type I TRDs, loop 2 is followed by an α helix. This may suggest that structure of type I TRDs deviates from that of *Hha*I at this junction. We propose that our alignment indicates that the TRDs of type I S subunits contain a DNA sequence recognition region consisting of β strand 2 and loops 1 and 2 with the same tertiary structure as part of the TRD of *Hha*I C5-cytosine mtase.

The only other type II mtase structure cocrystallised with DNA is of the *Hae*III C5-cytosine mtase. The *Hae*III TRD is slightly less ordered than that of *Hha*I but the overall fold of the polypeptide backbone in the DNA recognition region is the same (Reinisch *et al*, 1995). Although all biochemically characterised type I R/M systems methylate the N6 position of adenine and the *Hha*I and *Hae*III mtases methylate the C5 position of cytosine, there is no

reason why they cannot use the same protein structure to recognise their DNA target since it has been shown for a number of mtases that the nucleotide which is the target for methylation is not a major determinant of sequence specificity (Klimasauskas & Roberts, 1995; Yang *et al*, 1995) and that sequence recognition can tolerate unusual base pairs (Smith *et al*, 1991). Our results make the experimentally testable prediction that amino acids in a well-defined and experimentally amenable region of the TRDs of type I S subunits are important for sequence recognition.

Discussion

Combining the methods of multiple sequence alignment and secondary structure prediction within the `sss_align` program has facilitated the alignment of all 51 known type I TRDs overcoming the difficulties imposed by the large size of the TRDs and their very limited sequence conservation. The alignment bears some similarity to a short section responsible for DNA sequence specificity in the *Hha*I mtase. We suggest that this implies that all TRDs of type I S subunits are the products of divergent evolution with a conserved tertiary structure and that part of this structure, by analogy with *Hha*I mtase, is involved in DNA sequence recognition.

A variety of experiments such as uv-induced crosslinking to DNA (Chen *et al*, 1995), chemical modification of lysines (Taylor *et al*, 1996), and random mutagenesis of TRDs (personal communication, M O'Neill and NE Murray) have been applied to the best characterised type I R/M systems, *Eco*KI and *Eco*R124I, to identify amino acids involved in sequence recognition. These experiments provide preliminary support for our identification of a DNA binding region.

Chemical modification of *Eco*R124I showed that several lysines in the second TRD were susceptible to modification especially in the absence of bound DNA (Taylor *et al*, 1996). Lysines 261,297 and 327 within the TRD were particularly strongly modified. Lys297 is the most strongly modified residue and lies within the second proposed recognition loop. These three lysine residues are also conserved in the first TRD of *Sty*SKI which recognises the same DNA

target as the second TRD of *Eco*R124I therefore supporting a role for them in sequence recognition (Thorpe *et al*, 1997). The other less strongly modified lysines in the second TRD may be required for non-specific DNA binding as they are not conserved in *Sty*SKI and lie outside of our predicted recognition region.

Random mutagenesis of the first TRD of *Eco*KI has so far changed 40 out of 150 amino acids (personal communication, M O'Neill and NE Murray). Most of the mutations are silent, but 3 of 5 mutations that impair restriction and modification are within the two putative recognition loops.

UV-crosslinking demonstrated that Tyr27 in the first TRD of *Eco*KI was in contact with the 3′ thymine base in the sequence complementary to the 5′AAC part of the *Eco*KI target (Chen *et al*, 1995). This residue is outside of our predicted recognition loops, however, it has been found that changing it to other amino acids has a minor effect on DNA specificity suggesting that it may be involved in a non- sequence specific interaction with the DNA (personal communication, M. O'Neill and N. E. Murray).

Genes similar to the *hsd* genes of enteric bacteria have now been found in non-enteric bacteria and archaebacteria (see references in table 1) indicating that type I R/M systems are widespread in nature. It has been suggested that diversity within genes such as those forming type I R/M systems would be advantageous to a bacterial population (Levin, 1988; Korona & Levin, 1993). Furthermore, the diversity in *hsd* gene sequences observed in enteric bacteria provides support for horizontal gene transfer and a very ancient origin for the *hsd* genes (Sharp *et al*, 1992; Murray *et al*, 1993). The presence of type I R/M systems on conjugative plasmids would assist the spread of *hsd* genes by horizontal transfer (Tyndall *et al*, 1997). The existence of a common tertiary structure for TRDs, as implied by figure 1, would support this model for the distribution of type I systems in nature. Gene duplication of TRDs and transfer of TRDs by recombination is evident, not only from genetic and sequencing experiments (Gough & Murray, 1983; Argos, 1985; Gann *et al*, 1987; Kannan *et al*, 1989; Dybvig & Yu, 1994), but also from biochemical results on the domain structure of the S subunit (Abadjieva *et al*, 1993;

Meister *et al*, 1993; Cooper & Dryden, 1994; Webb *et al*, 1995). Recombination was responsible for the generation of two new type I target specificities, *Sty*SQI and *Eco*R124/3I (Fuller-Pace *et al*, 1984; Fuller-Pace & Murray, 1986; Price *et al*, 1989), and evidence for recombination of a short stretch of the *hsdS* gene between *E.coli* B and *S.enterica* serovar *Potsdam* has been found (Sharp *et al*, 1992). It is possible that other recombination events could encompass the short region within the TRD which we have predicted to be involved in DNA recognition, thereby allowing the generation of new specificities. These experiments suggest that the type I S subunit is a fusion of two half S subunits each containing one TRD to give a two-fold rotationally symmetric arrangement of the TRDs and a bipartite DNA target (Cooper & Dryden, 1994; Kneale, 1994; Dryden *et al*, 1995). Horizontal gene transfer has also been proposed for the type II R/M systems (Jeltsch & Pingoud, 1996). The range of organisms in which type I systems have been found or postulated, and their diversity within species such as *E.coli* and *S.enterica*, could also suggest that a large pool of TRDs existed before the evolution of different bacterial species. Therefore, it may be possible to have similar TRDs in different species even without invoking horizontal transfer, if they both carried with them the same range of TRDs when the species diverged (Maynard-Smith *et al*, 1993).

If our alignments are realistic, then the similarity between TRDs of type I N6-adenine mtases and the TRDs of C5-cytosine mtases may extend to many, if not all, TRDs of type II N6-adenine mtases, type III mtases and other mtases which do not fit current classifications. This would support and extend the proposal (Wilson & Murray, 1991) that all mtases have evolved from a common ancestor consisting of a small monomeric TRD, such as that still found in *Aqu*I mtase (Karreman & de Waard, 1990) and *Eco*HK31I mtase (Lee *et al*, 1995), associated with a separate catalytic subunit. It has been proposed that the mtase catalytic subunit may have developed from early DNA repair enzymes which use the same base flipping method to gain access to their target base as the mtases (Roberts, 1995). The normal rate of mutation and gene duplication events coupled with the selection pressure within a bacterial population to expand the range of DNA target sequences, has virtually obscured this common origin. A conserved tertiary structure within TRDs implies that it may eventually be feasible to derive

the amino acid recognition code used by TRDs to recognise DNA sequences as is currently being revealed for zinc finger-DNA recognition (Choo & Klug, 1996).

Materials and methods

Most nucleotide or amino acid sequences of the S subunits were obtained from published references or GenBank and a database constructed which separated the sequences into TRDs and conserved spacer regions. The amino acid sequences for the S subunits of *Bsu*CI and *Kpn*AI were generously provided by Prof. T. Trautner and Dr. G. Xu (Berlin), and Dr. J. Ryu (Loma Linda). The locations of the TRDs in the S subunit sequence and, if known, their DNA target and type I family are given in table 1. The amino terminal TRD and carboxyl terminal TRD are indicated by the suffix -1 or -2 appended to the type I system's name in figure 1.

A database of TRDs was made by separating conserved and unconserved regions of the S subunits and discarding the conserved regions. This database was inverted (every member was compared with every other one) using `sss_align`, a new implementation of the Smith/Waterman algorithm (Smith & Waterman, 1981) using the Dayhoff PAM scoring scheme (Dayhoff *et al*, 1978). Each TRD sequence, along with its closest homologues, was then sent to the `PHD` program (Rost & Sander, 1993; Rost *et al*, 1994) and a secondary structure prediction acquired. Each prediction was then placed in a new database and again inverted using `sss_align` but this time including the secondary structure prediction as well as amino acid sequence. In this instance, `sss_align` performs better than normal sequence alignment methods in aligning two distantly related sequences because the addition of secondary structure information, whether derived from a real structure or a prediction as in this case, is used to help the alignment pass through regions of very low sequence identity. The output of `sss_align` was again used to cluster sequences of high similarity and overlap these clusters with others until nearly all the TRD sequences were successfully aligned. Some TRD sequences could not be inserted into this alignment by the program due to a lack of obvious homology and these were

aligned manually. These sequences are indicated in figure 1 by an asterisk after the TRD name. In addition, `sss_align` also aligned the known tertiary structure of *Hha*I mtase (Klimasauskas *et al*, 1994) with the *Eco*KI-1 TRD to provide a key for the prediction of the location of loops and strands involved in DNA recognition.

`sss_align` can be accessed at `http://www.icmb.ed.ac.uk/sss_align/`. Using secondary structure information from known structures, `sss_align` has been shown to successfully align sequences with only 15% amino acid identity (A.Coulson, pers. comm., CASP2, Second meeting on the critical assessment of techniques for protein structure prediction on World Wide Web URL: `http://iris4.carb.nist.gov/casp2/`). `sss_align` also adjusts for the variation in the reliability of the secondary structure predictions by using the residue by residue reliability of PHD predictions. This allows the program to align sequences even if parts have incorrectly predicted secondary structure.

Acknowledgements

We wish to thank Professor Noreen Murray, Dr. Andrew Coulson and our colleagues in their laboratories, particularly Dr. Mary O'Neill, for provision of unpublished data and many useful discussions. We also thank Professor Thomas Trautner and Dr. Guoliang Xu (Berlin), and Dr. Junichi Ryu (Loma Linda) for the provision of unpublished sequences and other information. This work would not have been possible without the support of The Royal Society and The Darwin Trust. David Dryden thanks the Royal Society for a University Research Fellowship and Shane Sturrock thanks the Biochemical and Biological Sciences Research Council for a studentship.

References

Argos, P. (1985) Evidence for a repeating domain in type I restriction enzymes. *EMBO J.*, **4**, 1351–1355.

Abadjieva, A., Patel, J., Webb, M., Zinkevich, V. & Firman, K. (1993) A deletion mutant of the type IC restriction endonuclease *Eco*R124I expressing a novel DNA specificity. *Nucl. Acids Res.*, **21**, 4435–4443.

Barcus, V. A. & Murray, N. E. (1995) Barriers to recombination: restriction. In *Population Genetics of Bacteria.*, (ed. S. Baumberg, J.P.W. Young, S.R. Saunders & E.M.H. Saunders.) Society for General Microbiology Symposium **52**. p 31–58.

Barcus, V.A., Titheradge, A.J.B. & Murray, N.E. (1995) The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli*. *Genetics*, **140**, 1187–1197.

Bickle, T.A. & Kruger, D.H. (1993) Biology of DNA restriction. *Microbiol. Rev.*, **57**, 434–450.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J-F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S.M., Weidman, J.F., Fuhrmann, J.L., Nguyen, D., Utterback, T.R., Kelley, J.M., Peterson, J.D., Sadow, P.W., Hanna, M.C., Cotton, M.D., Roberts, K.M., Hurst, M.A., Kaine, B.P., Borodovsky, M., Klenk, H-P., Fraser, C.M., Smith, H.O., Woese, C.R. & Venter, J.C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.

Chen, A., Powell, L.M., Dryden, D.T.F., Murray, N.E. & Brown, T. (1995) Tyrosine 27 of the specificity polypeptide of *Eco*KI can be UV crosslinked to a bromodeoxyuridine-substituted DNA target sequence. *Nucl. Acids Res.*, **23**, 1177–1183.

Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W. & Roberts, R. (1993) Crystal structure of the *Hha*I DNA methyltransferase complexed with S-adenosyl-L-methionine. *Cell*, **74**, 299–307.

Cheng, X. & Blumenthal, R.M. (1996) Finding a basis for flipping bases. *Curr. Biol.*, **4**, 639–645.

Choo, Y. & Klug, A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opinion in Struct. Biol.*, **7**, 117–125.

Cooper, L.P. & Dryden, D.T.F. (1994) The domains of a type I DNA methyltransferase: interactions and role in recognition of DNA methylation. *J. Mol. Biol.*, **236**, 1011–1021.

Cowan, G.M., Gann, A.A.F. & Murray, N.E. (1989) Conservation of complex DNA recognition domains between families of restriction enzymes. *Cell*, **56**, 103–109.

Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) Atlas of protein sequence and structure, vol. **5**, suppl. 3, p345–352. National Biomedical Research Foundation, Washington DC.

Dryden, D. T. F., Sturrock, S. S. & Winter, M. (1995) Structural modelling of a type I DNA methyltransferase. *Nature Struct. Biol.*, **2**, 632–635.

Dybvig, K. & Yu, H. (1994) Regulation of a restriction and modification system via DNA inversion in *Mycoplasma pulmonis*. *Molec. Microbiol.*, **12**, 547–560.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L-I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.F., Smith, H.O. & Venter, J.C. (1995) Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Fuller-Pace, F.V. & Murray, N.E. (1986) Two DNA recognition domains of the specificity polypeptides of a family of type I restriction enzymes. *Proc. Natl. Acad. Sci. USA*, **83**, 9368–9372.

Fuller-Pace, F.V., Bullas, L.R., Delius, H. & Murray, N.E. (1984) Genetic recombination can generate altered restriction specificity. *Proc. Natl. Acad. Sci. USA*, **81**, 6095–6099.

Gann, A.A.F., Campbell, A.J.B., Collins, J.F., Coulson, A.F.W. & Murray, N.E. (1987) Reassortment of DNA recognition domains and the evolution of new specificities. *Molec. Microbiol.*, **1**, 13–22.

Gough, J.A. & Murray, N.E. (1983) Sequence diversity among related genes for recognition of specific targets in DNA molecules. *J. Mol. Biol.*, **166**, 1–19.

Gubler,M., Braguglia, D., Meyer, J., Piekarowicz, A. & Bickle, T.A. (1992) Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction-modification enzymes. *EMBO J.*, **11**, 233–240.

Harrison, S.C. (1991) A structural taxonomy of DNA-binding domains. *Nature*, **353**, 715–719.

Jeltsch, A. & Pingoud, A. (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.*, **42**, 91–96.

Kan, N.C., Lautenberger, J.A., Edgell, M.H. & Hutchison III, C.A. (1979) The nucleotide sequence recognized by the *Escherichia coli* K12 restriction and modification enzymes. *J. Mol. Biol.*, **130**, 191–209.

Kannan, P., Cowan, G.M., Daniel, A.S., Gann, A.A.F. & Murray, N.E. (1989) Conservation of organisation in the specificity polypeptides of two families of type I restriction enzymes. *J. Mol. Biol.*, **209**, 335–344.

Karreman, C. & de Waard, A. (1990) *Agmenellum quadruplicatum* M.*Aqu*I, a novel modification methylase. *J. Bacteriol.*, **172**, 266–272.

King, G. & Murray, N. E. (1994) Restriction enzymes in cells, not eppendorfs. *Trends in Microbiol.*, **2**, 465–469.

Klimasauskas, S., Kumar, S., Roberts, R. & Cheng, X. (1994) *Hha*I methyltransferase flips its target base out of the DNA helix. *Cell*, **76**, 357–369.

Klimasauskas, S. & Roberts, R. (1995) M.*Hha*I binds tightly to substrates containing mismatches at the target base. *Nucl. Acids Res.*, **23**, 1388–1395.

Kneale, G.G. (1994) A symmetrical model for the domain structure of type I DNA methyltransferases *J. Mol. Biol.*, **243**, 1–5.

Korona, R. & Levin, B.R. (1993) Phage-mediated selection and the evolution and maintenance of restriction-modification. *Evolution*, **47**, 556–575.

Kroger, M. & Hobom, G. (1984) The nucleotide sequence recognised by the *Escherichia coli* A restriction and modification enzyme. *Nucl. Acids Res.*, **12**, 887–899.

Labahn, J., Granzin, J., Schluckebier, G., Robinson, D.P., Jack, W.E., Schildkraut, I. & Saenger, W. (1994) Three dimensional structure of the adenine specific DNA methyltransferase M.*Taq*I in complex with the cofactor S-adenosylmethionine. *Proc. Natl. Acad. Sci. USA*, **91**, 10957–10961.

Lange, C., Wild, C. & Trautner, T.A. (1996) Identification of a subdomain within DNA-(cytosine-C5)-methyltransferases responsible for the recognition of the 5′ part of their DNA target. *EMBO J.*, **15**, 1443–1450.

Lautenberger, J.A., Kan, N.C., Lackey, D., Linn, S., Edgell, M.H. & Huthchison III, C.A. (1978) Recognition site of *Escherichia coli* B restriction enzyme on ϕXsB1 and simian virus 40 DNAs: an interupted sequence. *Proc. Natl. Acad. Sci. USA*, **75**, 2271–2275.

Lee, K-F., Kam, K-M. & Shaw, P-C. (1995) A bacterial methyltransferase M.*Eco*HK31I requires two proteins for in vitro methylation. *Nucl. Acids Res.*, **223**, 103–108.

Levin, B.R. (1988) Frequency dependent selection in bacterial populations. *Phil. Trans. Roy. Soc. London, Ser. B*, **319**, 459–472.

Luisi, B. (1995) DNA-protein interaction at high resolution. In *DNA-protein: structural interactions.*, (ed. D.M.J. Lilley) Oxford University Press, p 1–48.

Malone, T., Blumenthal, R.M. & Cheng, X. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyl-transferases, and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.*, **253**, 618–632.

Maynard-Smith, J., Smith, N.H., O'rourke, M. & Spratt, B.G. (1993) How clonal are bacteria? *Proc. Natl. Acad. Sci. USA*, **90**, 4384–4388.

Meister, J., MacWilliams, M., Hubner, P., Jutte, H., Skrzypek, E., Piekarowicz, A. & Bickle, T.A. (1993) Macroevolution by transposition: drastic modification of DNA recognition

by a type I restriction enzyme following Tn5 transposition. *EMBO J.*, **12**, 4585–4591.

Murray, N.E., Daniel, A.S., Cowan, G.M. & Sharp, P.M. (1993) Conservation of motifs within the unusually variable polypeptide sequences of type I restrcition and modification enzymes. *Molec. Microbiol.*, **9**, 133–143.

Nagaraja, V., Steiger, M., Nager, C., Hadi, S.M. & Bickle, T.A. (1985a) The nucleotide sequence recognised by the *Escherichia coli* D type I restriction and modification system. *Nucl. Acids. Res.*, **13**, 389–399.

Nagaraja, V., Shepherd, J.C.W., Pripfl, T. & Bickle, T.A. (1985b) Two type I restriction enzymes from *Salmonella* species: purification and DNA recognition sequences. *J. Mol. Biol.*, **182**, 579–587.

Price, C., Shepherd, J.C.W. & Bickle, T.A. (1987) DNA recognition by a new family of type I restriction enzymes: a unique relationship between two different DNA specificities. *EMBO J.*, **6**, 1493–1497.

Price, C., Lingner, J., Bickle, T.A., Firman, K. & Glover, S.W. (1989) Basis for changes in DNA recognition by the *Eco*R124 and *Eco*R124/3 type I DNA restriction and modification enzymes. *J. Mol. Biol.*, **205**, 115–125.

Ravetch, J.V., Horiuchi, K. & Zinder, N.D. (1978) Nucleotide sequence of the recognition site for the restriction-modification enzyme of *Escherichia coli* B. *Proc. Natl. Acad. Sci. USA*, **75**, 2266–2270.

Reinisch, K.M., Chen, L., Verdine, G.L. & Lipscomb, W.N. (1995) The crystal structure of *Hae*III methyltransferase covalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. *Cell*, **82**, 143–153.

Roberts, R.J. (1995) On base flipping. Cell **82**, 9–12. Rost, B. & Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost, B., Sander, C. & Schneider, R. (1994) PHD — an automatic mail server for protein secondary structure prediction. *CABIOS*, **10**, 53–60.

Schluckebier, G., Labahn, J., Granzin, J., Schildkraut, I. & Saenger, W. (1995) A model for DNA binding and enzyme action derived from crystallographic studies of the *Taq*I N6-adenine-methyltransferase. *Gene*, **157**, 131–134.

Sharp, P.M., Kelleher, J.E., Daniel, A.S., Cowan, G.M. & Murray, N.E. (1992) Roles of selection and recombination in the evolution of type I restriction-modification systems in enterobacteria. *Proc. Natl. Acad. Sci. USA*, **89**, 9836–9840.

Smith, S.S., Kan, J.L.C., Baker, D.J., Kaplan, B.E. & Dembek, P. (1991) Recognition of unusual DNA structures by human DNA(cytosine-5)methyltransferase. *J. Mol. Biol.*, **217**, 39–51.

Smith, T.F. & Waterman M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Somer, R. & Schaller, H. (1979) Nucleotide sequence of the recognition site of the B-specific restriction modification system of *Escherichia coli*. *Mol. Gen. Genet.*, **168**, 331–335.

Suri, B., Shepherd, J.C.W. & Bickle, T.A. (1984) The EcoA restriction and modification system of *Escherichia coli* $15T^-$: enzyme structure and DNA recognition sequence. *EMBO J.*, **3**, 575–579.

Taylor, I.A., Webb, M. & Kneale, G.G. (1996) Surface labelling of the type I methyltransferase M.*Eco*R124I reveals lysine residues critical for DNA binding. *J. Mol. Biol.*, **258**, 62–73.

Titheradge, A.J.B., Ternent, D. & Murray, N.E. (1996) A third family of allelic *hsd* genes in *Salmonella enterica*: sequence comparisons with related proteins identify conserved regions implicated in restriction of DNA. *Molec. Microbiol.*, **22**, 437–447.

Thorpe, P.H. (1995) The DNA specificity of type I restriction and modification enzymes. Ph.D. Thesis, University of Edinburgh.

Thorpe, P.H., Ternent, D. & Murray, N.E. (1997) The specificity of *Sty*SKI, a type I restriction enzyme, implies a structure with rotational symmetry. *Nucl. Acids Res.*, **25**, 1694–1700.

Tyndall, C., Meister, J. & Bickle, T.A. (1994) The *Escherichia coli prr* region encodes a functional type IC DNA restriction system closely integrated with an anticodon nuclease

gene. *J. Mol. Biol.*, **237**, 266–274.

Tyndall, C., Lehnerr, H., Sandmeier, U., Kulik, E. & Bickle, T.A. (1997) The type IC *hsd* loci of enterobacteria are flanked by DNA with high homology to the phage P1 genome: implications for the evolution and spread of DNA restriction systems. *Molec. Microbiol.*, **23**, 729–736.

Valinluck, B., Lee, N.S. & Ryu, J. (1995) A new restriction-modification system, *Kpn*BI, recognized in *Klebsiella pneumoniae*. *Gene*, **167**, 59–62.

Webb, M., Taylor, I.A., Firman, K. & Kneale, G.G. (1995) Probing the domain structure of the type IC DNA methyltransferase M.*Eco*R124I by limited proteolysis. *J. Mol.Biol.*, **250**, 181–190.

Wilke, K., Rauhut, E., Noyer-Weidner, M., Lauster, R., Pawlwek, B., Behrens, B. & Trautner, T.A. (1988) Sequential order of target-recognising domains in multispecific DNA-methyltransferases. *EMBO J.*, **7**, 2601–2609.

Wilson, G.G. & Murray, N.E. (1991) Restriction and modification systems. *Ann. Rev. Genet.*, **25**, 585–627.

Xu, G., Willert, J., Kapfer, W. & Trautner, T.A. (1995) *Bsu*CI, a type I restriction-modification system in *Bacillus subtilis*. *Gene*, **157**, 59.

Yang, A.S., Shen, J-C., Zingg, J-M., Mi, S. & Jones, P.A. (1995) *Hha*I and *Hpa*II DNA methyltransferases bind DNA mismatches, methylate uracil and block DNA repair. *Nucl. Acids Res.*, **23**, 1380–1387.

Figure legends

Figure 1. Multiple sequence alignments and predicted secondary structures for 51 TRDs from type I S subunits. For each TRD, the amino acids are coloured red, blue, green and pink for acidic, basic, polar and hydrophobic side chains. β strands are shown as yellow arrows and α helices as red rectangles. The sequence and known secondary structure of the TRD of *Hha*I type II mtase are shown at the top of the diagram followed by the type I TRDs arranged such that they are closest relatives as determined by `sss_align`. β strands 1, 2 and 3, and loops 1 and 2 which are involved in DNA recognition by *Hha*I and predicted to have the same role in type I TRDs, are marked above the *Hha*I secondary structure. Strands 1 and 2 comprise amino acids 228–231 and 240–243 in *Hha*I respectively. In the absence of DNA, amino acids 250–253 in *Hha*I form a β strand.

Figure 2. Part of the TRD of *Hha*I mtase bound to its DNA target with loops 1 and 2 highlighted in magenta and cyan respectively (Klimasauskas *et al*, 1994). β strands 1, 2 and 3 are shown in red, green and yellow respectively. The extrahelical cytosine base projects out of the minor groove into the catalytic site within the methylation domain. This methylation domain is present in the M subunits of type I R/M enzymes (Dryden *et al*, 1995).

Footnotes for Table 1

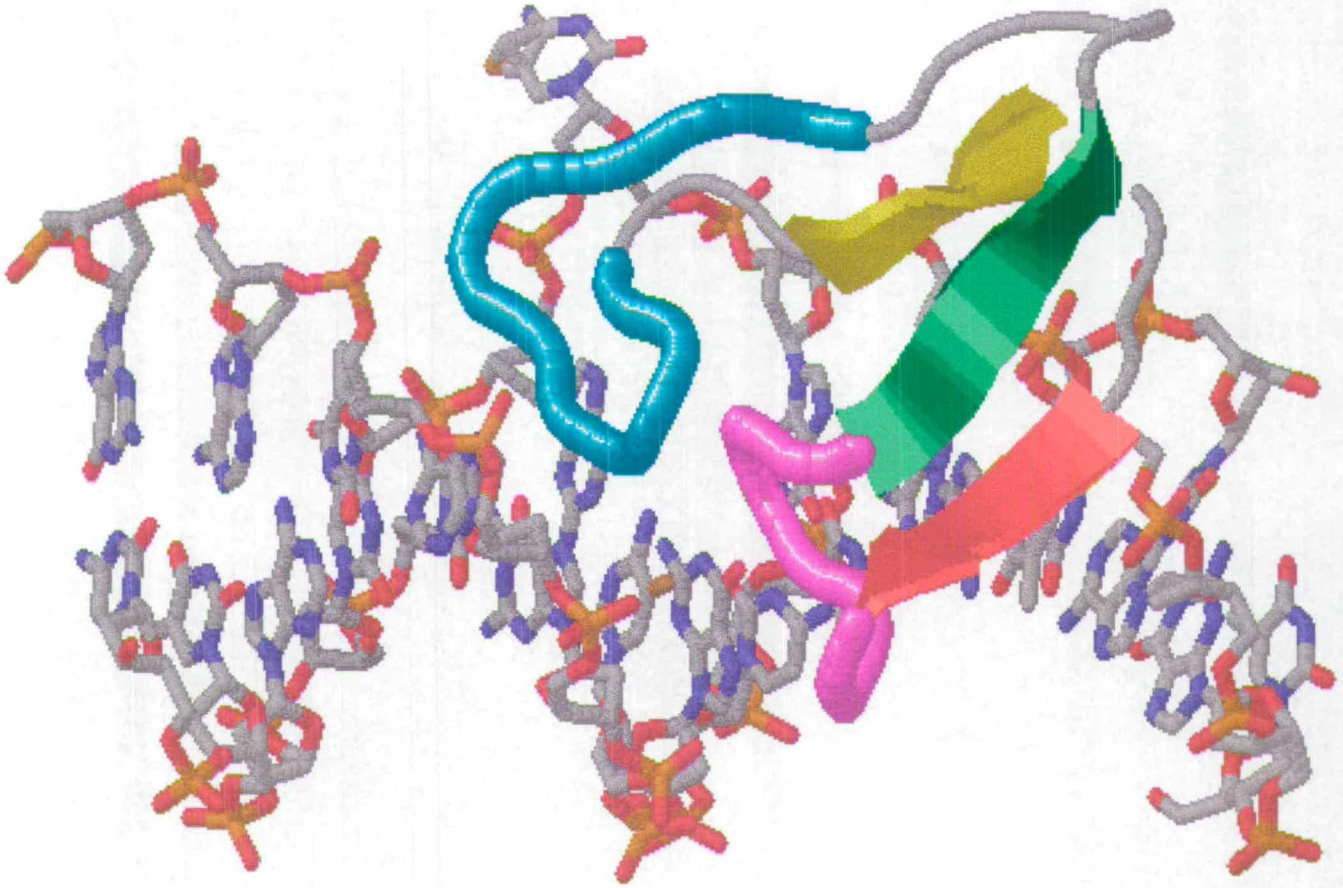
a. *Eco*, *Esherichia coli*; *Sty*, *Salmonella enterica*; *Cfr*, *Citrobacter freundii*; *Bsu*, *Bacillus subtilis*; *Kpn*, *Klebsiella pneunomiae*; *Mpu*, *Mycoplasma pulmonis*; HI, *Haemophilus influenzae* gene number; mj, *Methanococcus jannaschi* gene number.

b. Where the target sequence is known, the reference for this work is given.

Family	Name[a]	DNA target if known	S subunit length	1st TRD approximate location	2nd TRD approximate location	Reference[b]
IA	*Eco*KI	AAC N6 GTGC	464	11–157	214–368	Kan *et al*, 1979
IA	*Eco*BI	TGA N8 TGCT	474	11–158	215–379	Lautenberger *et al*, 1978; Ravetch *et al*, 1978; Somer & Schaller, 1979
IA	*Eco*DI	TTA N7 GTCY	444	11–128	185–348	Nagaraja *et al*, 1985a
IA	*Sty*LTIII	GAG N6 RTAYG	469	11–153	209–375	Nagaraja *et al*, 1985b
IA	*Sty*SPI	AAC N6 GTRC	463	11–157	214–367	*ibid.*
IA	*Eco*R5I		>140	1–140		Barcus *et al*, 1995; Thorpe, 1995
IA	*Eco*R10I		>131	1–131		*ibid.*
IA	*Eco*R13I		>152	1–152		*ibid.*
IB	*Eco*AI	GAG N7 GTCA	589	110–247	403–540	Kroger & Hobom, 1984; Suri *et al*, 1984
IB	*Eco*EI	GAG N7 ATGC	594	109–247	403–545	Cowan *et al*, 1989
IB	*Cfr*AI	GCA N8 GTGG	578	108–236	385–529	Kannan *et al*, 1989
IB	*Sty*SKI	CGAT N7 GTTA	587	108–249	396–538	Thorpe *et al*, 1997
IB	*Sty*STI		>146	1–146		Thorpe, 1995
IB	*Eco*R17I	ATR.....	>138	1–138		Barcus *et al*, 1995; Thorpe, 1995
IC	*Eco*R124I	GAA N6 RTCG	409	25–142	215–350	Price *et al*, 1987
IC	*Eco*DXXI	TCA N7 RTTC	406	23–139	211–341	Gubler *et al*, 1992
IC	*Eco*prrI	CCA N7 RTGC	405	22–159	232–360	Tyndall *et al*, 1994
ID	*Sty*SBLI	CGA N6 TACC	434	1–153	229–405	Titheradge *et al*, 1996
ID	*Eco*R9I		464	1–188	264–435	Barcus *et al*, 1995; N. Murray pers. comm.
IC?	*Bsu*CI	GAY N7 TGGA	405	23–162	219–355	Xu *et al*, 1995, G. Xu & T. Trautner pers. comm.
IC?	*Kpn*AI		439	1–155	231–410	Valinluck *et al*,1995, J Ryu pers. comm.
IC?	*Mpu*1AI		401	1–139	221–359	Dybvig & Yu, 1994
IC?	*Mpu*1BI		336	1–139	188–324	*ibid.*
ID?	HI0216		385	20–138	198–333	Fleischmann *et al*, 1995
ID?	HI1286		459	1–176	268–445	*ibid.*; Titheradge *et al*, 1996
?	mj0130		425	23–161	231–371	Bult *et al*, 1996
?	mj1218		425	28–155	226–368	*ibid.*
?	mj1531		425	28–170	241–371	*ibid.*

Table 1: S subunits of type I restriction/modification systems

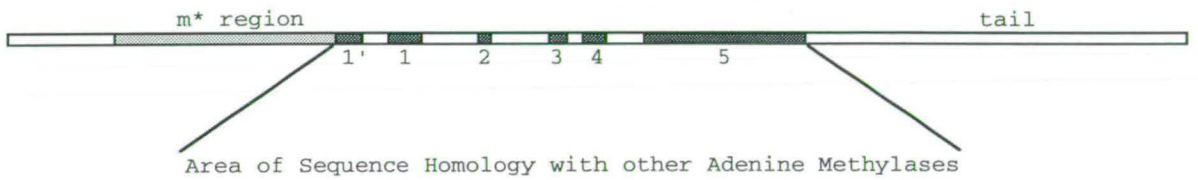

β1
L1
β2
L2
β3
Hha1_trd
EcoKI_1
StySPI_1
EcoR10I *
EcoR17I
EcoKI_2
EcoAI_1
EcoEI_1
StyLTIII_1
mj1531_1
StySPI_2
CfrAI_2
EcoEI_2
EcoR5I
EcoDXXI_2
BsuCI_2
HI0216_2
EcoR124I_1
CfrAI_1
mj1531_2
StySBLI_2
HI1286_2
KpnAI_2
BsuCI_1
mj0130_2
EcoprrI_1
mj0130_1
EcoR124I_2
StySKI_1
MpuIAI_2
mj1218_1
StySBLI_1
KpnAI_1
HI1286_1
EcoAI_2
mj1218_2
StySTI *
StySKI_2
EcoBI_1
EcoDI_2 *
EcoR9I_2
EcoBI_2
EcoR9I_1
StyLTIII_2
EcoDXXI_1
EcoDI_1
MpuIAI_1.MpuIBI_1
MpuIBI_2
HI0216_1 *
EcoR13I *
EcoprrI_2 *



Structural modelling of a type I DNA methyltransferase

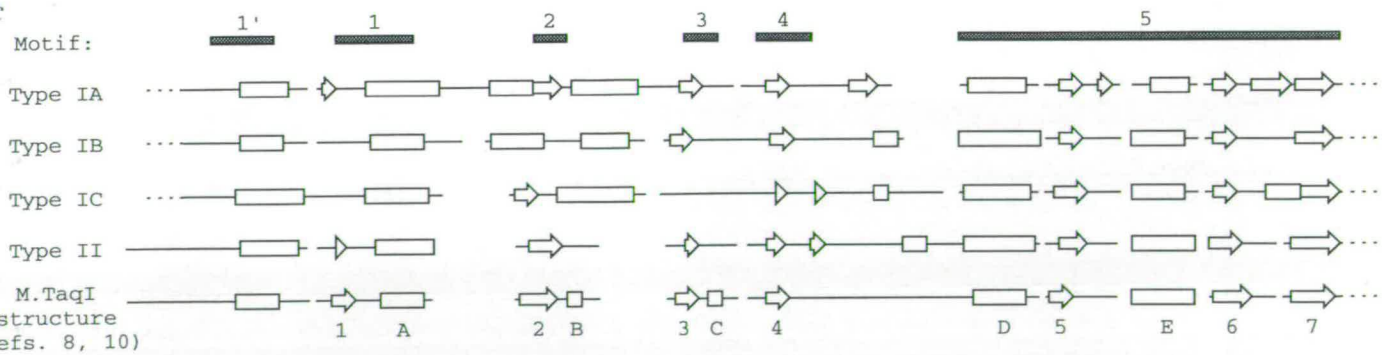
Amino-acid sequence comparison and tertiary structure modelling suggest a structure for type I DNA methyltransferases and an evolutionary link to type II DNA methyltransferases.

Sir—We have identified, by amino-acid sequence alignment and secondary structure prediction, a domain in the modification subunits of type I DNA restriction–modification (R–M) systems that is identical to that found in the simpler methyltransferases (mtases) of type II R–M systems and which is responsible for catalysing DNA methylation. Using the known structures of type II mtases, we have constructed a tertiary structure model of this domain, and a model of the entire mtase part of a type I system bound to DNA. The models correlate very well with all of the available data on type I systems and strongly suggest an evolutionary link between all mtases in type I and type II R–M systems.

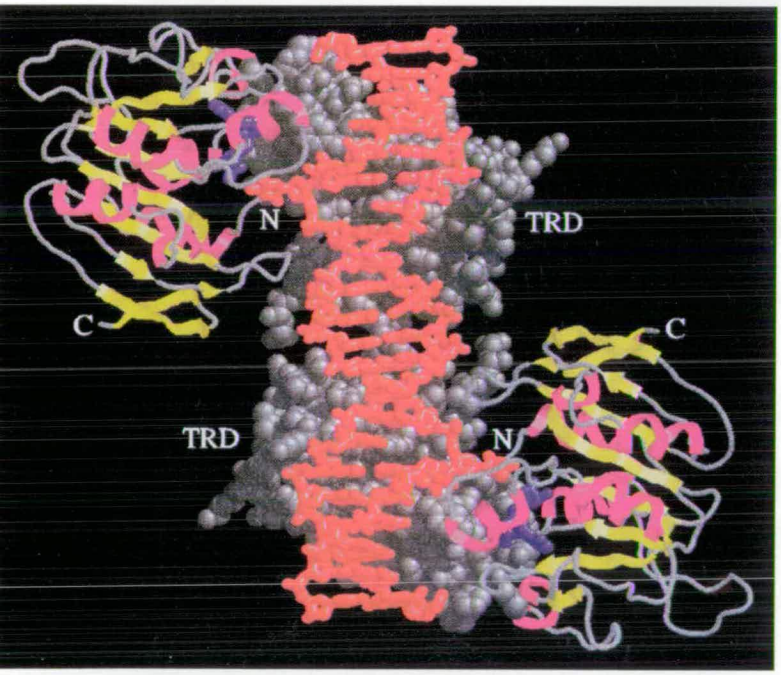



```
Motif                  1'                    1                    2          3          4
M.EcoKI       (146) GQYFTPRPLIKTI (11) VQDPAAGTAGFLIEA (26) FIGLELV (25) IRLGNTL (08) KAHIVATNPPF (17)
M.EcoAI       (161) GEFYTPRAVTRFM (11) IMDPACGTGGFLACA (21) IHGVEKK (21) IRHDNTL (09) QLDVIVTNPPF (19)
M.EcoR124I    (196) GEFFTPQHVSKLI (13) IYDPAAGSGSLLLQA (12) FFGQEIN (22) IKLGNTL (09) PFDAIVSNPPY (28)
M.Mycoplasma  (195) GEFYTPSKVSELL (13) AYDPACGSGSLLIKL (09) IYGQEVK (22) LRSGDTL (10) SFDCIVANPPF (26)
M.TaqI        ( 18) GRVETPPEVVDFM (11) VLEPACADGPFLRAF (09) FVGVEID (11) GILADFL (06) AFDLILGNPPY (32)

Motif                                                                5
M.EcoKI       (17) NKQLCFMQHIIETL HPGGRAAVVVPDNVLF  EGGKGTDIRRDLMDKCHLHTILRLPTGIFYA QGVKTNVLFFTK (165)
M.EcoAI       (19) ETADLFLQLIVEVL AKNGRAAVVLPDGTLF  GEGVKTKIKKLLTEECNLHTIVRLPNGVFNPYTGIKTNLLFFTK (120)
M.EcoR124I    (28) KADFAFVLHALNYL SAKGRAAIVCFPGIFY  RGGAEQKIRQYLVDNNYVETVISLAPNLFFG TTIAVNILVLSK (114)
M.Mycoplasma  (26) YADFAFLQHMLFHVNKDNGIIASVFSLGILSRKSPKAEEDIRKYIIDKNYIDTIIFLPPNLFYN TSIESCIIVARK (116)
M.TaqI        (32) NLYGAFLEKAVRLL KPGGVLVFVVPATWLV   LEDFALLREFLARE   KTSVYYLGEVFPQ KKVSAVVIRFQK (211)
```



ig. 1 *a*, The M subunit (529 amino acids) of *Eco*KI showing the N-terminal region containing the m* mutations which abolish the preference of *Eco*KI or hemimethylated DNA substrates (153 amino acids long)[3], the region of homology with the catalytic domain of type II mtases (200 amino acids) and he tail region (~ 150 amino acids)[4]. *b*, Sequence alignment showing six conserved blocks in the archetypal M-subunits of the three type I families, the M-subunit of *Mycoplasma pulmonis*[22], the first non-enteric bacterial type I R–M system and *Taq*I type II mtase, the archetypal member of the γ class of N6 adenine mtases[5]. Six additional M subunits and 14 type II γ class mtases were also included in the alignment but are not shown since they are almost identical to the archetypal M-subunits[1,2] or their alignment is similar to that previously published[12]. The numbers in brackets refer to the number of amino acids not shown and which are not in the conserved blocks. *c*, Secondary structure predictions[13], for the region in (*b*) above for the type I families and the type II γ class of mtases, compared with the secondary structure observed in the catalytic domain of the *Taq*I mtase crystal structure[8]. Arrows and rectangles indicate β-strands and α-helices respectively. The strands and helices in the *Taq*I structure are numbered as in ref. 10.



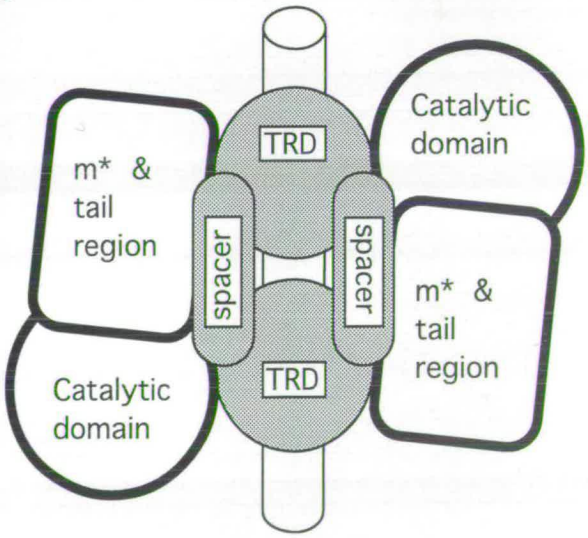


Fig 2. *a*, A tertiary structure model, constructed using FRODO[23] and SYBYL (Tripos) software, of the catalytic domain from amino acids 141 to 380 in the M-subunit of *Eco*KI based on the sequence alignment in Fig. 1 and the structure of the *Taq*I mtase[8]. The α-helices, β-strands and SAM are coloured magenta, yellow and blue respectively. The helices and strands are numbered as given in Fig. 1*c*. The sites of mutagenesis and proteolysis are coloured red and green respectively and numbered as in Table 1. *b*, A front view of a partial model of a type I mtase bound to DNA constructed using two copies of the structure of the type II *Hha*I mtase bound to DNA[7]. The catalytic domain of *Hha*I has been replaced with our modelled domain with helices, strands and SAM coloured as in (*a*). The target bases are flipped out of the DNA helix, shown in red, to project into the catalytic domain and are 10 base pairs apart as found in the *Eco*KI recognition sequence. The grey space-filled structures are the two TRDs proposed to form the S-subunit. The actual structure of these domains in type I mtases is unknown so the diagram shows the TRD of *Hha*I. They are located in two successive major grooves in agreement with DNA footprinting and protein–DNA crosslinking experiments[18,19]. *c*, A schematic rear view of the complete mtase structure bound to DNA. The M-subunits are shown in a bold outline with the postulated position of the N-terminal m* and C-terminal tail regions and the catalytic domain indicated. The S-subunit is shown shaded with the TRD joined by the two short spacer regions which are well conserved in all S-subunits. The DNA helix, shown as a cylinder, lies on the S-subunit and between the M-subunits in agreement with earlier models[17,21].

The function of a bacterial DNA R–M system is to maintain the methylation state of the chromosome by methylating the hemimethylated DNA produced during replication, and restrict viral propagation by cleaving unmodified viral DNA[1,2]. In a type I R–M system, the restriction endonuclease, modification mtase and sequence recognition functions require different subunits (R, M and S) in one large multifunctional enzyme[1,2]. The subunit stoichiometry is believed to be $R_2M_2S_1$, but the M_2S_1 form can function alone as a mtase of M_r 170,000. The type I sytems have been divided into families of related enzymes. The S subunits contain two DNA target recognition domains (TRDs) which each recognise one part of the bipartite targets of type I systems. The target for *Eco*KI, one of the best characterised type I systems, is A<u>A</u>C(N_6)G<u>T</u>GC. The TRDs are joined by two short sequences which are conserved in all S-subunits. The M-subunits methylate the adenine nucleotide, one on each strand, at each of the underlined positions in the sequence above, using S-adenosyl methionine (SAM) as cofactor. The M subunit of *Eco*KI contains a N-terminal m* region and a C-terminal tail region which together give *Eco*KI its preference for methylating

Table 1 Sites of proteolysis and mutation in the putative catalytic domain of *Eco*KI

Proteolytic cleavage sites in catalytic domain[4]	Comment
R168 ↓ E169	Trypsin cleavage at partially exposed loop, fragment from amino acid 169 to ~ 440. On β-strand 1.
V263 ↓ A264 A264 ↓ T265 R305 ↓ A306	Elastase or trypsin cleavage at buried site on β-strand. Fragments from cleavage site to end of M-subunit. On β-strands 4 and 5.
R279 ↓ T280 F281 ↓ V282	Trypsin or chymotrypsin cleavage at exposed loop. Fragments from cleavage site to end of M-subunit. SAM binding prevents cleavage and stabilises whole subunit. On loop between β-strand 4 and α-helix D.
Site-directed mutations[15]	
G177D	Abolishes SAM binding, insoluble when expressed at 37°C, soluble at 25°C. Mtase structure and DNA binding normal. In motif 1 between β-strand 1 and α-helix A.
N266D, F269G, F269C	SAM and DNA binding normal but inactive. In motif 4 at end of β-strand 4.
F269Y, F269W	SAM and DNA binding normal. Activity 25 % and ≤5 % of wild-type mtase respectively.

hemimethylated targets rather than unmodified ones[3,4] (Fig. 1*a*). The R subunit cleaves DNA at an apparently random site remote from the target sequence after extensive DNA translocation driven by ATP hydrolysis[1,2].

The type II R–M systems comprise separate endonuclease and mtase enzymes[1,2,5,6]. The tertiary structure of *Hha*I C5 cytosine mtase and *Taq*I N6 adenine mtase[7,8] shows that this class of mtases, with molecular masses between 20,000–40,000, generally contain only one TRD and a highly conserved catalytic domain which binds the SAM (X.Cheng & W. Saenger, personal communication)[9,10]. The conservation of sequence motifs and structural domains strongly suggests evolution from a common ancestor[10,11]. In contrast to the type I systems, the type II mtases generally recognise short, symmetric DNA targets and methylate unmodified DNA as efficiently as hemimethylated DNA.

We have compared the sequences of all the M-subunits of type I systems with the sequences of the γ class of type II N6 adenine mtases[5] (Fig 1*b*). This identified the same six sequence motifs in the M-subunits as found in the type II γ class mtases[12]. Secondary structure predictions over the 200 amino-acid region containing the six conserved blocks show α-helix/β-strand/α-helix repeats over the whole region (Fig. 1*c*). The structural features predicted by the PredictProtein PHD program[13] line up well with those actually found in the crystal structure of the *Taq*I mtase catalytic domain[8]. The insertions and deletions in the type I sequences,shown in Fig. 1*c*, occur on surface loops in the *Taq*I structure. This alignment of sequence and secondary structure suggests that the M-subunits of type I systems contain a catalytic domain virtually identical to that in *Taq*I. Using our sequence alignments and the *Taq*I tertiary structure, we have modelled the tertiary structure of the *Eco*KI M-subunit catalytic domain (Fig 2*a*). The model domain comprises two independent subassemblies joined by a loop between β-strands 3 and 4. The first subassembly has the form of the Rossman mononucleotide-binding fold[14]. The second subassembly is similar except for antiparallel β-strands 6 and 7, near the end of the domain. The model domain indicates plausible locations for the sites of mutagenesis[15] and proteolysis[4] which have been examined in *Eco*KI (Table 1). One can see that Gly 177 and Asn 266 are located close to the cofactor and that Phe 269 is poised at the edge of the active site where it can interact with the target base if it is flipped out of the DNA helix. Proteolysis yields stable polypeptides from the cleavage sites to the C terminus of the M-subunit[4]. Proteases can only reach the sites buried on the β-sheet when the N-terminal half of the subunit has been digested and the SAM-binding site lost. The exposed loop containing the Arg 279–Val 282 site is protected by SAM binding from digestion, suggesting a conformational change possibly analogous to that observed for the similar loop in *Hha*I mtase[7].

Following the recent suggestions that a type I mtase has approximate two fold rotational symmetry[4,16,17] we have constructed a model of the entire mtase part of a type I system bound to DNA using two copies of the type II *Hha*I mtase structure (Fig. 2*b*).The location of the start of the modelled domain, which contains two m* mutation sites at Ser 144 and Arg 153 (ref. 3) suggests that the N-terminal m* region is folded up against the back of the catalytic domain. The tail of the M-subunit is sensitive to the methylation state of the DNA target[4]. To allow one catalytic domain, bound at one half of the DNA target, to be sensitive to the methylation state of the other half of the target, it appears necessary to place the tail of each M-subunit against the DNA and TRD forming the second half of the target site (Fig. 2*c*). In the absence of further structural information, the m* region and the tail region are shown as a single block for each M-subunit in Fig. 2*c*.

In addition to being consistent with biochemical results, including recent DNA footprinting[18] and protein–DNA crosslinking[19] experiments, the model is also consistent with measurements of the hydrodynamic shape of the protein[17,20]. The effect of DNA binding to the *Eco*R124I type I mtase analysed by small angle X-ray scattering is particularly striking and has been interpreted as a clamping of the DNA between the M-subunits[17] exactly as shown in Fig. 2*b, c*. The proposed contacts between the M-subunits and both parts of the DNA target sequence could, by altering the sub-

unit orientation in response to target methylation, give *Eco*KI its specificity for methylating half-methylated DNA and cleaving unmodified DNA by the whole enzyme, as originally predicted by Burckhardt *et al*[21].

The apparent success of the modelling procedure also strongly suggests an evolutionary link between all mtases in type I and type II R–M systems. Almost all of the known R–M systems are of the type II class which contain a simple modification mtase, but recent results suggest that the more complex type I R-M systems may also be widespread in nature[2,22]. The link between type I and II mtases may be extended to suggest that all of the DNA mtases are constructed from one or more copies of the catalytic domain and one or more TRDs. These domains can be on the same polypeptide chain or on separate subunits. In some DNA mtases—the C5 cytosine[6], N6 adenine β-class[5] and type III N6 adenine mtases[1,2]—the catalytic domain is interrupted by the TRD. This could be accommodated structurally due to the nature of the two substructures within the catalytic domain and could arise by gene duplication and rearrangement. The presence of additional domains and subunits in type I and type III R–M systems compared to the type II R–M systems allows the acquisition of a more sophisticated range of enzymatic responses on binding to a DNA target.

David T.F. Dryden, Shane S. Sturrock and Markus Winter
Institute of Cell and Molecular Biology, University of Edinburgh, Darwin Building, The King's Buildings, Mayfield Road, Edinburgh EH9 3JR, UK

Correspondence should be addressed to D.T.F.D.

Received 27 April; accepted 20 June, 1995.

1. Wilson, G.G. & Murray, N.E. *Annu. Rev. Genet.* **25**, 585–627 (1991).
2. Barcus, V.A. & Murray, N.E. In *Population genetics of bacteria* in Society for General Microbiology Symposium **52** (eds. Baumberg, S., Young, J.P.W., Saunders, S.R. & Wellington, E.M.H.) 31–58 (Cambridge University Press, Cambridge; 1995).
3. Kelleher, J.E., Daniel, A.S. & Murray, N.E. *J. molec. Biol.* **221**, 431–440 (1991).
4. Cooper, L.P. & Dryden, D.T.F. *J. molec. Biol.* **236**, 1011–1021 (1994).
5. Wilson, G.G. *Meth. Enzym.* **216**, 259–279 (1992).
6. Kumar, S. *et al. Nucleic Acids Res.* **22**, 1–10 (1994).
7. Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. *Cell* **76**, 357-369 (1994).
8. Labahn, J. *et al. Proc. natn. Acad. Sci. U.S.A.* **91**, 10957–10961 (1994).
9. Cheng, X. *Annu. Rev. Biophys. Biomol. Struct.* **24**, 293–318 (1995).
10. Schluckebier, G., O'Gara, M., Saenger, W. & Cheng, X. *J. molec. Biol.* **247**, 16–20 (1995).
11. Lauster, R. *J. molec. Biol.* **206**, 313–321 (1989).
12. Noyer-Weidner, M., Walter, J., Terschüren, P.-A., Chai, S. & Trautner, T. A. *Nucleic Acids Res.* **22**, 4066–4072 (1994).
13. Rost, B. & Sander, C. *J. molec. Biol.* **232**, 584–599 (1993).
14. Rossman, M.G., Liljas, A., Bränden, C.-I. & Banaszak, L. J. in *The Enzymes, 3rd ed. vol XI.* (ed.Boyer, P.D.) 61–102 (Academic Press ; 1975).
15. Willcock, D.F., Dryden, D.T.F. & Murray, N.E. *EMBO J.* **13**, 3902–3908 (1994).
16. Meister, J. *et al. EMBO J.* **12**, 4585–4591 (1993)
17. Taylor, I.A., Davis, K.G., Watts, D. & Kneale, G.G. *EMBO J.* **13**, 5772–5778 (1994).
18. Powell, L.M. & Murray, N.E. *Nucleic Acids Res.* **23**, 967–974 (1995).
19. Chen, A., Powell, L.M., Dryden, D.T.F., Murray, N.E. & Brown, T. *Nucleic Acids Res.* **23**, 1177–1183 (1995).
20. Powell, L. M., Dryden, D. T. F., Willcock, D. F., Pain, R. H. & Murray, N. E. *J. Mol Biol.* **234**, 60-71 (1993).
21. Burckhardt, J., Weisemann, J., Hamilton, D. L. & Yuan, R. *J. molec. Biol.* **153**, 425–440 (1981).
22. Dybvig, K. & Yu, H. *Molec. Microbiol.* **12**, 547–560 (1994).
23. Jones, T.A. *J. appl. Crystallogr.* **11**, 268–272 (1978).